University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Matija Piškorec

STATISTICAL INFERENCE OF EXOGENOUS AND ENDOGENOUS INFORMATION PROPAGATION IN SOCIAL NETWORKS

DOCTORAL THESIS

Zagreb, 2019

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Matija Piškorec

STATISTICAL INFERENCE OF EXOGENOUS AND ENDOGENOUS INFORMATION PROPAGATION IN SOCIAL NETWORKS

DOCTORAL THESIS

Supervisors:
Professor Mile Šikić, PhD
Tomislav Šmuc, PhD

Zagreb, 2019

Sveučilište u Zagrebu
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Matija Piškorec

STATISTIČKO ZAKLJUČIVANJE O EGZOGENOME I ENDOGENOME ŠIRENJU INFORMACIJA U DRUŠTVENIM MREŽAMA

DOKTORSKI RAD

Mentori:
prof. dr. sc. Mile Šikić
dr. sc. Tomislav Šmuc

Zagreb, 2019.

The dissertation was made at the University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Signal Processing, in cooperation with Ruđer Bošković Institute.

Supervisors: Professor Mile Šikić, PhD
Tomislav Šmuc, PhD

The dissertation has: 111 pages

The dissertation number:

About the Supervisors

Mile Šikić Ph.D. is a Professor at the University of Zagreb Faculty of Electrical Engineering and Computing, Croatia. At the same university, he obtained a Ph.D. in computer science in 2008. Currently he is spending sabbatical year at Genome Institute of Singapore. For the first 7 years of his career he worked as system integrator, consultant and project manager on projects with industry in the fields of computer and mobile networks. In 2009 he became Assistant Professor in computer science. His scientific work is focused on the development of new algorithms and machine learning methods for genome sequence analysis and analysis of dynamics in networks.

Tomislav Šmuc Ph.D. is a Head of Laboratory for Machine Learning and Knowledge Representation at Ruđer Bošković Institute, Zagreb. His research interest last twenty years is in the area of artificial intelligence, in development and use of machine learning and data mining techniques for knowledge discovery in different domains of science and technology. In this period he has been participating in, or leading, a number of research projects financed by Croatian, European and other international funding agencies. Tomislav Šmuc was mentor of a dozen of master's and Phd students at University of Zagreb; involved in organization of international conferences (ECML-PKDD, Discovery Science) on several occasions. Tomislav Šmuc has published over 100 papers in journals and proceedings of international conferences. He serves as a reviewer for a number of scientific journals in the fields of computer science, computational biology and interdisciplinary science, and is an evaluator for several research funding agencies.

O mentorima

Mile Šikić, dr. sc., profesor je na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu. Na istom je sveučilištu stekao titulu doktora znanosti iz računarstva 2008. godine. Trenutno provodi studijsku godinu na Genome Institute of Singapore. Prvih sedam godina svoje karijere radio je kao sistem integrator, konzultant i projektni rukovoditelj na projektima iz industrije iz područja računalnih i mobilnih mreža. 2009. postaje docent iz računarstva. Njegov znanstveni rad fokusiran je na razvoj novih algoritama i metoda strojnog učenja za analizu genomskih sekvenci i analizu dinamike u mrežama.

Tomislav Šmuc, dr. sc., voditelj je Laboratorija za strojno učenje i reprezentaciju znanja na Institutu Ruđer Bošković u Zagrebu. Njegovi istraživački interesi u zadnjih dvadeset godina su iz područja umjetne inteligencije, u razvoju i primjeni strojnog učenja i tehnika rudarenja podataka za otkrivanje znanja u različitim domenama znanosti i tehnologije. U tom je periodu sudjelovao, ili vodio, nekoliko istraživačkih projekata financiranih iz hrvatskih, europskih i ostalih međunarodnih agencija. Tomislav Šmuc je mentorirao na desetke diplomskih i doktorskih studenata na Sveučilištu u Zagrebu, a i sudjelovao je u organizaciju nekoliko međunarodnih konferencija (ECML-PKDD, Discovery Science). Tomislav Šmuc je objavio preko 100 članaka u časopisima i zbornicima međunarodnih konferencija. Sudjeluje kao recenzent za nekoliko znanstvenih časopisa iz područja računarstva, računalne biologije i interdisciplinarnih znanosti, i evaluator je za nekoliko istraživačkih agencija za financiranje.

Acknowledgements

It is impossible to thank everyone who, at one time or another, helped me directly or indirectly in my long PhD journey. So there will be omissions, for which I apologize in advance. Every person mentioned is here because of its own merit. Every omission is my own fault.

First, without digressing too much (I guess there is no word limit for the Acknowledgment section!), it would be wrong of me not to mention where it all started. The initial idea for the research that led to this thesis was not mine. In November 2013 my colleague at the Rudjer Boskovic Institute (RBI) Nino Antulov-Fantulin and my past and future advisor (although he did not know it at the time) Mile Šikić from the Faculty of electrical engineering and computing (FER), University of Zagreb, came to me with an idea to conduct an online survey on the upcoming referendum using Facebook API. The production version was finished in what essentially was a week-long informal hackaton by a handful of people - Bruno Rahle, Vedran Ivanac, Iva Miholić, Matej Mihelčić, Tomislav Lipić, together with Nino and myself. Nino, Mile and I immediately started considering research directions that we can take with this dataset, a process which took a while and involved several other people, most importantly Vinko Zlatić, from RBI, and Sebastian Krausse who was at that time a postdoc in Vinko's group. The most interesting pattern evident in data - the surges of user registrations after news media reports on our survey application, suggested that there is an interplay between peer and external influence (what we would later call endogenous and exogenous influence). Existing methods for estimation of these two influences from data either postulated too much information on the external sources of influence, or required more data on the activity of users, while in our case we had only a single activation cascade. What started as a side project eventually grew into a PhD thesis topic after I formally enrolled on a PhD program at FER in 2014 under dual supervision of Mile Šikić and Tomislav Šmuc from RBI.

I was lucky to be able to work on my doctoral research as a research assistant at RBI since 2013. Thanks to the funding from various projects we had at that time, I never had to worry about not being able to travel to conferences or workshops, or do all other things we scientist are expected to do but cost money. For this I am grateful to senior scientists at RBI, in particular my advisor Tomislav Šmuc. Probably the most important part of funding, my PhD tuition on FER, was funded from a Croatian Science Agency fund through a project led by Dragan Gamberger on RBI. Lastly, when my research assistant contract expired, I was employed as a young researcher at the Centre of Excellence for Data Science and Advanced Cooperative Systems, led by Sven Lončarić on FER.

Over the years on RBI I had an opportunity to work with many amazing people. I already mentioned Nino with whom I shared a room at RBI and with whom I had many

fruitful discussions, but I should also mention my other two roommates - Matej Mihelčić and Ilona Kulikovskikh. Colleagues and friends from the Division of Electronics - Maria Brbić, Damir Korenčić, Dijana Tolić, Dario Sitnik, Davor Oršolić, Ozren Jović, Vedrana Vidulin. Colleagues and friends from other parts of RBI - Boris Okorn, Tomislav Lipić, Davor Davidović, Ivan Sović, Eva Cetinić, Ivan Grubišić, Jelena Repar, Jelena Čubrić, Enis Afgan, Andrea Kadović, Irena Barjašić, Lea Liović, Ivan Ivek, Peter Škoda. Members of our informal weekly programming group at RBI: Vedran Ivanac, Željko Rumenjak, Goran Kopčak. These guys did not pay my rent, but they made my stay at RBI well worth it. Of the seniors with whom my research work would not be possible - Tomislav Šmuc, Vinko Zlatić, Sebastian Krausse and Hrvoje Štefančić. Seniors from the Division of Electronics - Dragan Gamberger, Ivan Marić, Darko Kolarić, Bono Lučić, Branka Medved-Rogina, Strahil Ristov, Ivica Kopriva.

Along with the people from RBI, I should also mention international collaborators with whom I had a pleasure to have formal or informal interaction over the years, or spend some time at their groups. Bojan Žagrović and members of his group from Max F. Perutz Laboratories in Vienna. Aristides Gionis and members of his Data Mining group at Aalto University, Helsinki. Dirk Helbing and members of his Computational Social Science group at ETH Zurich. Irena Vodenska from Boston University. Pretty much everyone from the Department of Knowledge Technologies at the Jozef Stefan Institute, Ljubljana.

Last but not the least, a special thanks to my family and friends. They will forgive me for not mentioning them by name, as they are far too numerous for that.

Abstract

In the last decade we witnessed a rapid rise of the online social media services. Although they were created in the early 2000's, their rise began in earnest after 2010 when their presence started to fundamentally alter the traditional media landscape. Today, their influence on the way our society consumes, curates and disseminates information is indisputable. With their wider adoption came also the first criticism, as well as a need to solve emerging legislative, ethical and societal issues. One line of research is to explain and quantify the sources of influence in online social services and investigate to what extent are these new social landscapes vulnerable to manipulation by third parties. This manipulation is often performed by using user's digital traces - a record of their activities on the online social service. These digital footprints have a potential to characterize users in more detail than what they themselves would be willing to share otherwise. For example, user's personality traits can be inferred indirectly from the content with which they interact through online services, and even their writing style on the written content they published could be used to infer their demographic characteristics. This opens opportunities for micro-targeting of users for various dubious purposes, for example by increasing their propensity to spread misinformation.

Research described in this thesis shows that much can be learned about user engagement by using very little data - in our case only friendship connections between users and a single activation cascade. A single activation cascade means we only have one registration event per user. This data alone is sufficient to estimate, under certain assumptions, whether activation for each user was predominantly influenced by its peers with which they are connected (endogenous influence), or the exogenous factors which are external to the friendship network itself. Both endogenous and exogenous factors, for example mass media, are known to have a significant impact on the activity of users of online social media.

The methodology developed in this thesis requires postulating an explicit endogenous influence model which governs interactions between pairs of users, while exogenous influence is assumed to act equally towards all users in the network. Several suitable endogenous influence models are proposed for the use with this methodology. First one is Susceptible-Infected model, commonly used in epidemiological modeling. Second one features a decay factor for the endogenous influence, which is a realistic assumption for in social systems. Third one features a logistic threshold for activation. Exogenous influence is modelled as an independent probability of activation which is, at any given time, equal for all non-activated users, although it may change in time.

An inference method is developed where maximum likelihood estimation is used to estimate relative magnitudes of endogenous and exogenous influence on users. These esti-

mates can then be used to characterize influence of individual users. The computational scalability analysis is performed on simulated data to demonstrate that the inference method is able to scale to large social networks.

Empirical data on over 20 thousand Facebook users is used for evaluation of the proposed inference method. Data is collected using three unique Facebook political survey applications which provided Facebook friendship relations between users and a single activation cascade - a single registration event per user. Referral links, which identify user's origin, are used as a proxy for user's activation type. Users whose referral links originated from Facebook are considered as endogenously activated while those whose referral links originated from an external website are considered as exogenously activated.

Inference method is used to estimate the most probable source of influence for each user individually, as well as to asses the overall influence of different media channels (peer communication, Facebook advertisements, or external news media) on user's activations cascade. Ethical, methodological and technical issues regarding data collection in the context of online social media services is discussed. Guidelines on how to collect online social media data in an ethically principled way are provided, especially in the context of satisfying requirements for reproducible research.

Estimating endogenous and exogenous influence in networks with a statistical methodology that is conceptually simple, yet powerful and efficient, is widely applicable to scientific domains where deciphering properties of spreading processes and external influences on complex networks is crucial for an explanation of new phenomena.

Keywords: online social networks, social influence estimation, statistical learning, maximum likelihood method, online social data collection

Statističko zaključivanje o egzogenom i endogenom širenju informacija u društvenim mrežama

Zadnjih deset godina svjedoci smo naglog uzleta popularsnosti online društvenih mreža. Iako postoje od ranih 2000-tih, njihov uspon je ozbiljno započeo tek nakon 2010. kada njihova prisutnost počinje fundamentalno mijenjati tradicionalne medije. Utjecaj online društvenih mreža na način na koji naše društvo konzumira, odabire i diseminira informacije je danas neporeciv. S njihovom širom upotrebom pojavile su se i prve kritike, kao i potreba za rješavanjem novonastalih legislativnih, etičkih i društvenih pitanja. Jedan smjer istraživanja pokušava objasniti i kvantificirati izvore utjecaja u online društvenim servisima i istražiti do koje mjere su oni podložni manipulaciji od treće strane. Ta manipulacija se često provodi korištenjem korisničkih digitalnih tragova - zapisa njihovih aktivnosti na online društvenim servisima. Navedeni digitalni otisci imaju potencijal za karakterizaciju korisnika s više detalja nego što su oni sami voljni otkriti. Primjerice, korisničke crte osobnosti i demografske karakteristike se mogu procjeniti indirektno preko sadržaja ili stila pisanja kojeg korisnici koriste na online servisu. Ovo otvara mogućnost za mikro-ciljanje (eng. micro-targeting) korisnika u svrhu različitih sumnjivih radnji ili propagande, primjerice povećavanjem njihove sklonosti da šire dezinformacije.

Istraživanje opisano u ovoj disertaciji pokazuje da se mnogo toga može saznati o aktivnosti korisnika koristeći relativno malo podataka - u našem slučaju riječ je samo o podacima o prijateljskim vezama između korisnika i jednoj kaskadi širenja informacija, pri čemu informacija koja se širi odgovara činu registracije (aktivacije) korisnika na online društvenom servisu. Koristeći samo ove podatke moguće je, pod određenim pretpostavkama, zaključiti je li aktivacija svakog pojedinog korisnika pretežno uzrokovana zbog njegovih prijatelja s kojima su povezani (endogeni utjecaj) ili faktorima van društvene mreže (egzogeni utjecaj). Poznato je da i endogeni i egzogeni faktori, primjerice iz medija, imaju značajan utjecaj na aktivnost korisnika.

U Poglavlju 1 opisana je motivacija i pregled područja istraživanja iz širenja informacija u online društvenim mrežama, kao i statističkih metoda koje se koriste prilikom modeliranja širenja informacija iz empirijskih podataka. Opisani su ciljevi doktorskog istraživanja koji se sastoje od definiranja modela endogenog i egzogenog širenja informacija u društvenim mrežama, razvoja metode za statističko zaključivanje parametara navedenih modela iz podataka, i evaluacije navedene metode na empirijskim podacima prikupljenih iz stvarnih online društvenih mreža.

U Poglavlju 2 opisani su modeli širenja informacija koji se koriste u metodi statističkog zaključivanja razvijenoj u sklopu ovog doktorskog istraživanja. Metoda zahtjeva postuliranje izričitog modela endogenog utjecaja koji definira interakcije između parova korisnika.

S druge strane, pretpostavka kod egzogenog utjecaja je da djeluje jednako prema svim korisnicima u društvenoj mreži. Predloženo je nekoliko primjerenih modela endogenog utjecaja koji se mogu koristiti u tu svrhu. Prvi je Susceptible-Infected model, često korišten u epidemiološkom modeliranju, gdje svaki trenutno aktivni korisnik ima nezavisnu priliku aktivirati bilo kojeg od svojih prijatelja u online društvenoj mreži, pri čemu se vjerojatnost aktivacije ne mijenja u vremenu. Drugi model pretpostavlja eksponencijalno opadajući utjecaj što znači da tijekom vremena korisnici imaju sve manju vjerojatnost aktivirati nekog od svojih prijatelja, što je realistična pretpostavka u društvenim interakcijama. U trećem modelu se vjerojatnost aktivacije mijenja s brojem prethodno aktiviranih prijatelja prema logističkoj funkciji, što znači da postoji prag broja prethodno aktiviranih prijatelja koji se mora dostići prije nego vjerojatnost aktivacije dostigne značajnu vrijednost. Egzogeni utjecaj je modeliran kao nezavisna vjerojatnost aktivacije koja je, u svakom danom trenutku, jednaka za sve još neaktivne korisnike, iako se može mijenjati u vremenu.

Modeli endogenog i egzogenog utjecaja objedinjeni su unutar funkcije izglednosti (eng. likelihood) koja daje vjerojatnost svake kombinacije parametara modela, uvjetno s obzirom na promatrane podatke koji se u ovom slučaju sastoje od mreže prijateljstva između korisnika i vremena njihove aktivacije. U Poglavlju 3 opisana je razvijena metoda statističkog zaključivanja koja koristi maksimalnu izglednost (eng. maximum likelihood) za pronalaženje parametara endogenog i egzogenog utjecaja. Ti parameteri se potom koriste za procjenu relativne magnitude endogenog i egzogenog utjecaja na korisnika pomoću mjere *egzogene odgovornosti* (eng. exogenous responsibility) koja na skali od 0 do 1 kvantificira koliko je na korisnikovu aktivaciju utjecao egzogeni utjecaj, pri čemu veća vrijednost označava jači egzogeni utjecaj. Definiraju se i mjere *individualnog i kolektivnog utjecaja* (eng. individual and collective influence) koje kvantificiraju utjecaj pojedinog korisnika i grupe korisnika na aktivacije njihovih prijatelja u društvenoj mreži, pri čemu se uzima u obzir samo endogena komponenta utjecaja.

Metoda statističkog zaključivanja koristi metodu maksimalne izglednosti za procjenu fiksnog skupa parametera endogenog utjecaja koji su isti za sve korisnike i ne mijenju se u vremenu. S druge strane, egzogeni utjecaj se procjenjuje u svakom vremenskom trenutku zasebno pa broj parametara ovisi o broju diskretnih vremenskih trenutaka. U realnim primjenama gdje se zahtjeva određena vremenska granulacija egzogenog utjecaja to uvijek rezultira prevelikim brojem parametara za izravnu procjenu metodom maksimalne izglednosti. Zbog toga je razvijena alternirajuća optimizacijska metoda gdje se parametri endgenog i egzogenog utjecaja naizmjence fiksiraju kako bi se smanjio broj parametara koji se optimiraju u svakoj iteraciji algoritma. Manji broj parametara omogućuje da se optimizacija provede nekom od standardnih metoda numeričke optimizacije. Iako ne postoji teorijska garancija konvergencije metode, praksa pokazuje da je za konvergenciju svih

parametara potrebno svega nekoliko iteracija algoritma. Provedena je analiza računske skalabilnosti kako bi se pokazalo da predložena alternirajuća metoda statističkog zaključivanja skalira čak i na velike društvene mreže od preko 20 tisuća korisnika.

Evaluacija je prvo provedena na simuliranim podacima pri čemu su aktivacijske kaskade korisnika simulirane prema jednom od tri predložena modela endogenog utjecaja. Egzogeni utjecaj dizajniran je tako da sadrži nekoliko distinktnih eksponencijalno-opadajućih šiljaka u vremenu. Ovo je obrazac koji se često opaža u empirijskim podacima, primjerice kad medijske objave uzrokuju porast interesa i pojačanu aktivaciju korisnika. Predložena metoda statističkog zaključivanja sposobna je precizno odrediti stvarne parametre endogenog i egzogenog utjecaja u simuliranom slučaju, kao i stvarni razlog aktivacije svakog pojedinog korisnika, koristeći samo podatke o mreži prijateljstava između korisnika i vrijeme aktivacije svakog pojedinog korisnika. Provedeni su opsežni eksperimenti na simuliranim podacima gdje je pokazano da metoda dobro radi i na proizvoljnim krivuljama egzogenog utjecaja. Također, rezultati su uspoređeni s onima dobivenima jednostavnom osnovnom (eng. baseline) metodom gdje su svi korisnici koji u trenutku aktivacije nisu imali drugih aktiviranih prijatelja proglašeni egzogeno aktiviranima. Ova jednostavna metoda podcjenjuje stvarni broj egzogeno aktiviranih korisnika, pogotovo pred kraj aktivacijske kaskade kada je većina korisnika u mreži već aktivirana. Zbog specifičnog načina prikupljanja podataka o korisnicima - korisnici koji čine mrežu prijateljstava su svi oni koji se u konačnici aktiviraju, mreža prijateljstava se pred kraj aktivacijske kaskade zasiti s aktiviranim korisnicima što ne odražava stvarno stanje u društvenoj mreži. Ovaj efekt nazivamo *pristranost opažača* (eng. observer bias) i on uzrokuje precjenjivanje egzogenog utjecaja kako se približavamo kraju aktivacijske kaskade. Kako bi se on izbjegao u funkciju izglednosti dodan je korekcijski faktor.

U Poglavlju 2 opisana je metodologija prikupljanja podataka korištenih u empirijskoj evaluaciji. Za empirijsku evaluaciju su korišteni podaci o preko 20 tisuća korisnika društvene mreže Facebook. Podaci su prikupljeni pomoću tri online političke ankete koje koriste Facebook Graph programsko sučelje za registraciju korisnika. Ankete su provedene na hrvatskom jeziku i vezane su za tri različita politička događaja u Hrvatskoj - referendum o pitanju ustavne definicije braka iz 2013. i parlamentarne izbore 2015. i 2016. godine. Prikupljeni podaci sadrže informaciju o prijateljskim poveznicama između korisnika i samo jednu aktivacijsku kaskadu - vrijeme registracije svakog pojedinog korisnika. Referencijske poveznice (eng. referral links), koje identificiraju porijeklo korisnika, su korištene kao aproksimacija za korisnikov tip aktivacije. Korisnici čija je referencijska poveznica potekla s Facebooka su smatrani endogeno aktiviranima, dok su oni čija je referencijska poveznica potekla s vanjske web stranice smatrani egzogeno aktiviranima. Anketne aplikacije su bile aktivne otprilike tjedan dana prije samog dana glasanja i ti-

jekom tog vremena su privukle medijsku pozornost online novinskih portala koji su u svojim objavama dijelili poveznicu na aplikacije. U trenucima takvih objava vidljiv je skok u registraciji korisnika na anketne aplikacije što ukazuje na egzogeni utjecaj jer se korisnici registriraju na aplikaciju potaknuti vanjskim izvorom. S druge strane, struktura mreže prijatelja ukazuje na efekt homofilije - korisnici se pretežno povezuju s drugim korisnicima koji dijele njihove političke stavove, ili su im slični po nekim drugim karakteristikama (primjerice starosti), što ukazuje na endogeni utjecaj. Eksploratorna analiza prikupljenih podataka pokazuje da su strukturalne karakteristike mreže prijateljstava i statističke karakteristike demografije korisnika reprezentativne za hrvatski Facebook prostor. Raspravlja se i o etičkim, metodološkim i tehničkim aspektima prikupljanja podataka u kontekstu online društvenih mreža. Predstavljene su i smjernice za prikupljanje podataka s online društvenih mreža na etički prihvatljiv način, tako da se istovremeno poštuju privatnost korisnika, uvjeti korištenja online društvenih servisa kao i zahtjevi za reproducibilnost provedenog istraživanja.

Empirijska evaluacija predložene metode statističkog zaključivanja opisana je u Poglavlju 4. Pomoću prikupljenih empirijskih podataka procjenjuje se najvjerojatniji izvor utjecaja za svakog korisnika zasebno, kao i ukupni utjecaj svakog komunikacijskog kanala (komunikacija između korisnika, Facebook oglasi, vanjski medijski izvori) na korisničku aktivacijsku kaskadu. Kao metrika evaluacije koristi se površina ispod krivulje (eng. area under the curve - AUC) koja na empirijskim podacima postiže vrijednost od 0.7 do 0.8, što ukazuje na dobru diskriminacijsku moć predložene metode statističkog zaključivanja u kontekstu binarnog klasifikacijskog problema gdje se korisnici klasificiraju na endogeno i egzogeno aktivirane prema njihovim referencijskim poveznicama. Od komunikacijskih kanala kao najjutjecajnija se pokazala direktna komunikacija između korisnika, dok su se vanjski medijski izvori pokazali dominantni samo na jednom skupu podataka gdje udio egzogeno aktiviranih korisnici čine većinu (preko 90% od ukupnog broja korisnika). Provedena je i usporedba predložene mjere individualnog utjecaja svakog pojedinog korisnika sa strukturalnim mjerama izračunatima iz mreže prijateljstava, pri čemu je najjača korelacija s Pagerank centralnošću.

U sklopu ovog doktorskog istraživanja razvijena je metoda statističkog zaključivanja za procjenu endogenog i egzogenog širenja informacija u društvenim mrežama, no potencijalna primjena nadilazi primjenu u samo jednoj specifičnoj domeni. Identifikacija egzogenih utjecaja ima potencijalnu primjenu i u analizi financijskih sustava gdje vanjski utjecaji mogu imati ključnu ulogu u dinamici sustava. Također, paradigma identifikacije endogenog i egzogenog utjecaja potencijalno ima širu primjenu u modeliranju općenitih dinamičkih sustava gdje bi se pomoću takvih metoda identificirale ranjivosti sustava na vanjske šokove, kao i podložnost manipulaciji od trećih strana. Procjena endogenog i

egzogenog utjecaja u mrežama sa statističkom metodologijom koja je konceptualno jednostavna, a opet snažna i učinkovita, široko je primjenjiva u znanstvenim područjima gdje je dešifriranje svojstava procesa širenja i vanjskog utjecaja na kompleksnim mrežama ključno za objašnjavanje novih pojava.

Ključne riječi: online društvene mreže, procjena društvenog utjecaja, statističko zaključivanje, metoda maksimalne izglednosti, prikupljanje podataka s društvenih mreža

Contents

Chapter 1

Introduction

1.1 Motivation and related work

Growing popularity of *online social networking services* means that a large amount of human interaction is now recorded in digital form. An online social networking service is an information system which gives its users an ability to efficiently communicate with other users. However, unlike traditional online communication services like email, its purpose is ultimately to build online social communities which are connected through shared interests, and to ease and enhance interaction within them. This is done through easy identification of shared interests, codifying modes of communication (for example, through easy expression of approval or disapproval) and aggregating interaction in a way to give each user a glimpse into the community's consensus. Availability of this kind of data in digital form provides an opportunity to investigate social interactions on a scale that was previously unattainable [1, 2, 3, 4, 5, 6, 7, 8]. Most studied online social systems are blogspace [9] and online social networks services such as Flickr [10], Twitter [11], Facebook [12] and Instagram [13]. At the same time, the same availability of data on something that was previously in the personal domain raises ethical concerns which were previously not encountered [14, 15]. An interesting research question is to what extent is human interaction facilitated by information systems, and whether and how their potential can be misused.

The most general definition of *influence* between entities in complex systems is that it is a conditional dependence between entity's states [16]. If we have no other information except sequences of state for each of the entities, the most straightforward way to measure this conditional dependence statistically is with Hidden Markov Models (HMM's) [17]. An assumption of HMM's is the Markovian property - therefore depends only on the current state, regardless of the past.

In this work we are interested in *social influence* between users in online social net-

works. Here, a social influence is defined as the degree to which the behavior of individuals changes the behavior of their peers [18], and an online social network is mathematically understood as a set of all possible pairs of users through which influence could arise due to technological or social context. Mathematical modeling of social influence is an active field of research in sociology for decades [19, 20, 21], especially actor models which model conditions under which nodes in networks change their social connections [21]. These sociological studies used mostly methods from graph theory and agent based modeling. Recent surge of research into complex network structure and dynamics [22, 23] introduced new set of methods as well as new research communities to the problem of social influence. These new methods were mostly from statistical physics and computer science and were more suitable for addressing technical and methodological challenges inherent in the analysis of large online social datasets.

Underlying any type of social influence is a some kind of *social interaction* between persons in social network. For example, an act of transmitting a piece of information from one person to another could be understood as a social interaction. For our purpose this is particularly important as the digital communication technology facilitates the spread of information and allows it to be stored more efficiently and in form which is more accessible for subsequent analysis. A spread of information between multiple users of an information system is an *information cascade*.

Research on information cascades usually focuses on the prediction of future evolution of an information cascade given past diffusion traces [5]. These methods usually use some variant of Linear threshold model (LT) [19], Independent cascade model (IC) [24], Susceptible infected (SI) or Susceptible infected susceptible (SIS) models as the underlying model of information propagation. The benefit of these models is that they do not contain any hidden state, just the two observable states - active and inactive, which simplifies inference from data. LT model can be inferred with gradient ascent method [25], while asynchronous versions of LT and IC (AsLT and AsIC) can be inferred with maximum likelihood estimation [26]. In theory, these models require that a structure of a social network is known, although there are ways to use them even in cases where there is no explicit information on the underlying social network. In these cases some assumptions should be imposed in order to perform inference, for example that all persons have the same probability to adopt the information [27], or that individual influence functions follow a specific form [8]. Investigating information cascades can give us insights on social influence between users of an online social network, and is already widely used in domains such as viral marketing [27], behavior adoption [28] and epidemic spreading [29].

Social influence is often confounded with correlation effects such as homophily - a tendency of similar persons to interact with each other due to factors other than a direct

influence [30, 31]. One way to account for this is with randomization strategies [32], which should diminish true influence and leave correlation intact. Exogenous factors that are easy to measure and are suspected to mediate social influence could also be explicitly accounted for in the modeling. Examples include news media [33] and real-life events such as political unrest [1] and natural disasters [34]. In general, exogenous factors could be modeled either indirectly as anonymous uniform influence that acts on all nodes in the network [35] or directly in the form of "authorities" which exert their influence on the nodes in the network [36]. Also, the type and characteristics of the information content that is being transmitted could also be a significant mediator of the social influence [37].

A full probabilistic representation of the influence model could be achieved with a *likelihood function*, which gives a probability of observing any combination of parameters conditioned on the observed data. The combination of parameters for which likelihood is maximized are called *maximum likelihood* parameters. They could be found with standard optimization methods such as gradient ascent method [25] or Expectation-Maximization (EM) algorithm [38]. In the context of information cascades, it is not necessary to explicitly condition on the structure of social network, as one could analyze interaction dynamics with information-theoretic measures such as transfer entropy [39].

One of the most general versions of such likelihood-based approaches is an unified model of social influence [40] which is able to explicitly accommodate many social interaction mechanisms such as pairwise influence, local neighborhood effects, aggregate social behavior and exogenous factors. Many commonly used information diffusion models could be represented as a special case of this unified model of social influence, including Complex Contagion Model [41], Independent Cascade Model (IC) [24], The Generalized Threshold Model [24] and The Linear Friendship Model (LT) [20]. There is currently no proposed method to fit this unified model in its most general form to data, although there are multiple proposed methods of fitting more specific likelihood-based models to data, for example Independent Cascade Model [38] and asynchronous versions of Independent Cascade and Linear Threshold models [42].

Likelihood-based approaches also allow one to choose which of the several proposed social influence models has more support in data given the goodness of fit and expressiveness of the model itself. In theory, parsimonious models - the ones that give best explanation for data while being relatively simple, should be preferred. Model selection criteria used in more traditional statistical contexts are not always best suited for selecting among social influence models [43]. For example, information-theoretic measures such as Akaike Information Criteria (AIC) and Minimum Description Length (MDL) can usually be used only in cases where there are few parameters which correspond directly to the complexity of a model. This is satisfied if the social influence is parametrized with a small

set of basis functions [44]. Resampling techniques such as k-fold crossvalidation [45] have a limited application in the context of social influence due to the combinatorial complexity of the social network structure. This is why the most pragmatic way to evaluate social influence models is through their predictive and explanatory power [46], keeping in mind the amount of data and the characteristics of the phenomena we are trying to model.

In general, trying to find a "true model" might be impossible in domains pertaining to human behavior where processes underlying observables are complex and heterogeneous [47]. What is more, while investigating human behavior it is nearly impossible to control sufficiently for all possible confounding conditions, and we often have to contend with the observational data instead. So we should never fool ourselves that our model selection outputs anything close to the "true model", which could be arbitrarily complex and whose complexity could inhibit its explanatory power derived from inference on finite data, and probably could be easily outmatched by a much simpler (although wrong) model [48].

Another important aspect of social influence modeling is the existence of exogenous factors which confound with endogenous factors. Similarly as with correlational effects, this confounding could be hard to eliminate using observational data alone [30]. In ideal conditions the exogenous influence is negligible [49] but usually has to be explicitly accounted for [35, 50, 51]. Rather than being a nuisance, exogenous influence is often crucial in understanding the way social influence acts. This happens due to two reasons. First, there are usually multiple information channels though which information can propagate, and many of these channels are exogenous to the social network itself. Examples include news media, advertisements, and most forms of direct communication such as email, instant messengers, and even offline word-of-mouth communication. Second, exogenous factors can act as mediators of social influence, often by encouraging social engagement. For example, exogenous events such as political unrest [1, 52] and natural disasters [34] are often strong mediators of social influence. Exogenous factors themselves are often not directly observable in the online social network, but usually can be inferred from the available data. Research on exogenous factors in online social networks and its interplay with endogenous influences gains more and more importance as it becomes increasingly evident that these systems could be manipulated by various interest groups [53].

1.2 Objectives

The general aim of this research is a data-driven characterization of social influence in online social networks and how human interaction is facilitated by online information systems. This will be done by the development of methods for modeling and inference of

social influence that are able to describe both the endogenous influence between users of an online social network and the influence exogenous to the network itself. Additionally, methodology for collecting online social network data will be described in order to validate these methods on empirical data. The hypothesis is that the endogenous and exogenous influence can be modeled directly conditioned on a particular form of endogenous influence model, and that users in online social network could be characterized based on their susceptibility to one or another type of influence. The basic components needed to conduct this research are the following:

1. **Context** (Information system which facilitates interaction): Online social network.
2. **Agents** (Who is interacting through an information system): Users of an online social network.
3. **Interaction** (How do entities interact): Users communicate within social network, but there is also an exogenous influence acting on the users.
4. **Measurement** (What we actually measure and record): User registration at application and friendships in social network.
5. **Mode of influence** (What makes entities interact in the way they do): Users are influenced conditioned on a specific endogenous influence model while exogenous influence acts uniformly on all users.

This thesis will present research on the estimation of endogenous and exogenous influence between users in online social networks. In our case we have an *activation cascade* of user registrations which closely resembles information cascade where an actual information content is being transmitted between users. The basic requirement for inference is that we have data on a particular *activation* cascade in online social network and an explicit social network between users through which endogenous influence could act. As it will be demonstrated in the thesis, only a single activation cascade is needed for efficient inference. Assuming a particular form of endogenous and exogenous influence we can infer the parameters of influence and estimate their magnitudes on user and global level, as well as characterize activation of each user or groups of users as being dominantly endogenous- or exogenous-driven. Similar methods for estimating endogenous and exogenous influence exist in literature, for example peer and authority model [36] which, however, requires explicit modeling of *authorities* responsible for exogenous influence, while in our case this is not necessary. The social network structure is used directly for the inference rather than implicitly [8]. Also, there is no direct reliance on some sort of a network statistic such as degree distribution [54].

Chapter 3 describes the models which are used for modeling influence. Two approaches are presented, first where modeling is done indirectly by analyzing statistical properties of the endogenous and exogenous influence, and second where explicit models of endogenous

and exogenous influence are joined in the likelihood function.

Two methods are employed for estimation of endogenous and exogenous influence - 1) modeling exogenous influence directly and 2) joint modeling of endogenous and exogenous influence. For the second approach a likelihood function will be used which gives us a fully probabilistic description of the inference problem, as well as allows the usage of standard optimization methods for inference. Dimensionality of the likelihood function depends on the forms for endogenous and exogenous influence. Endogenous influence models are usually low-dimensional, and it makes sense to assume all users share equal parameters of the endogenous influence. On the other hand, exogenous influence could be parametrized by a suitable closed form, or evaluated non-parametrically at each time step. This makes the number of exogenous influence parameters dependent on the number of time steps, which could lead to high-dimensional model. However, there is a possibility to solve this with a form of expectation-maximization method where parameters are estimated by alternating expectation and maximization steps. A similar approach is to optimize just a subset of parameters in turn while holding the others fixed. While this procedure does not guarantee an optimal solution, in practice it yields near-optimal results to a optimization problem which would otherwise fail to converge. Instead of a direct numerical optimization we could also use a Markov Chain Monte Carlo (MCMC) sampling which gives samples from the likelihood function, allowing us to estimate confidence intervals on the parameters. Chapter 4 describes a maximum likelihood method for the inference of endogenous and exogenous influence.

Data from an actual online social network is used for the evaluation of the inference methodology. An online survey application is developed as a separate web page which uses Facebook Graph application programming interface (API) [55] for the authentication of users. In this case the series of user registrations could be viewed as the information cascade * because the information on the application is spreading between Facebook users.

Online survey applications were related to three distinct political events which happened in Croatia in the period from 2013 to 2016 (Figure 2.4): 1) referendum on the definition of marriage in 2013, 2) parliamentary elections in Croatia in 2015 and 3) parliamentary elections in Croatia in 2016. Some form of authentication is crucial for the application because we want to prevent multiple voting on the survey, and using official Facebook Graph API will allow us to access user data that would be unavailable if the custom authentication mechanism was used. Most importantly, it allows us to access Facebook friendship relations between registered users. The application should satisfy following methodological and ethical requirements:

*Sometimes this is referred to as an *activation cascade* instead of information cascade, because users are getting "activated" when they register on the application.

- Data collection should be performed via encrypted secure connection provided by the official Facebook Graph API and it should comply with its the privacy policy and terms of use [†].
- The users should be informed about the procedure of data collection, and they should give informed consent in advance.
- All private data used in this research should be anonymized. Users should not be in position to access other users data during the data collection process, although they may have an access to aggregated data for all users.
- The users should be familiar with the fact that the information collected will be used only for scientific purposes and that the anonymized as well as aggregated data will be made available to the scientific community.

Online social network datasets where users have to give an explicit consent to collect their data are usually small and sparse, and so researchers have to rely on simulated datasets in order to validate their models. The media and public interest related to the real-life political events helped to engage Facebook users and allowed the collection of large amount of data. Chapter 2 explains in detail how data was collected on Facebook users, along with the design of online political survey applications and the methodology of data collection. It also discusses ethical challenges inherent in collection of user data from online services and gives practical recommendations for future researchers.

The methodology for estimation of endogenous and exogenous influence is evaluated on both simulated activation cascades and actual activation cascades collected from the online survey application. In both cases the assumption is that we only have a friendship network between users and a *single* activation cascade - registration time for each user. Most other research relies on the availability of multiple activation cascades, which makes the method described in this thesis applicable to cases where there is little available data on user activations. For simulated activation cascades either an actual friendship network or a configuration model of it are used. Configuration model of a network preserves the degree sequence - number of Facebook friends each user has. The simulation follows and Independent Cascade model where each user has certain probability of activating each of its Facebook friends at each time step, conditioning on a particular endogenous influence model. Exogenous influence is designed as a non-parametric curve which closely resembles influence curves we observe in actual data. The output of the method is a single estimate for each user of its propensity of being influenced by either endogenous or exogenous influence. This allows us to use area under the curve (AUC) as an evaluation measure which tells us how well does the estimates classify users in these two categories. For sabor2015 and sabor2016 datasets there is an information on referral links from which users visited

[†]Facebook's privacy policy is available at `https://developers.facebook.com/policy/`.

the survey application, and so this is used as gold-standard labels in the calculation of the AUC score. Chapter 5 presents the results of the evaluation on the empirical datasets of over 20 thousand Facebook users which were collected though Facebook political survey applications. Estimates are then used for estimation of individual and collective influence of various groups of users.

The methodology gives an estimate to what extent was *each user* influenced due to endogenous or exogenous factors. We can assign to each user its share of the endogenous influence present in the social network, and calculate it recursively for each individual user (individual influence) or for groups of user (collective influence). In this case influence corresponds to the expected number of users that will be activated due to the endogenous influence of an individual user or groups of users in the next time step. This quantity can then be used to identify the most influential users, as well as a building block for influence maximization methods - identifying which groups of users have the most influence in social network or which individual users to target with incentivization strategies. Estimates that are obtained agree with the baselines obtained from raw data of referral links from which users visited the survey applications. Chapter 5.3 describes in detail the methodology for estimating collective influence and presents results on the empirical datasets.

The main contributions of this thesis are the following:

1. **Model of exogenous and endogenous information propagation in social networks.** A probabilistic model of influence in a social network is proposed that assumes a particular functional form of endogenous influence between users while the exogenous influence is non-parametric. The model can be easily extended to include additional information on the users or the type or characteristics of the influence.
2. **Method for estimation of parameters of the proposed model of information propagation in social networks.** An inference method is developed which uses only a *single* activation cascade and a social network of users to estimate relative magnitudes of endogenous and exogenous influence for each user individually.
3. **Evaluation of the proposed methodology on empirical data from social networks.** Inference methodology is applied on three empirical Facebook datasets of over 20 thousand users that participated in one of three online political survey applications. Besides characterizing to what extent is each user's activation driven due to endogenous and exogenous influence, estimates of collective influence of various groups of users are also provided.

Chapter 2

Data

We present challenges we encountered while designing several online political survey applications: 1) Referendum on the definition of marriage in Croatia in 2013 (10175 respondents), 2) Parliamentary elections in Croatia in 2015 and 2016 (6909 and 3818 respondents), 3) Local elections in Zagreb, Croatia's capitol, in 2017 (1666 respondents). Online surveys allowed us to reach many more respondents than it would be possible with traditional methods. We critically examine technical, methodological and ethical challenges we encountered during the design and execution of these surveys.

We developed our online survey applications as separate web pages where users could register with their Facebook accounts, cast their votes on the upcoming election and see statistics for their Facebook friends and all users. Upon registration, users had to comply with both Facebook's and our own privacy policy to allow us to retrieve their personal data. Also, they were able to share the link to application through Facebook which mimics snowball sampling. Surveys were active one week prior to actual pooling day and the attention of news media and general public helped us attract new users. However, each subsequent online survey attracted less and less respondents due to loss of novelty. Nevertheless, we still managed to sample representative Facebook population - distributions of the number of friends and other demographic characteristics are comparable to the whole Facebook population [56].

Three main challenges we encountered while collecting social network data from Facebook:

1. **Methodological** (Section 2.4). Can we get a representative sample of the population of interest? If not, can we at least correct our sample as to be more representative. If still not, can we at least estimate error/uncertainty we are introducing into our estimate?
2. **Technical** (Sections 2.2 and 2.3). How to design online survey application so that it satisfies technical requirement of data collection. How to use APIs and technology

to collect necessary data. And how to incentivize users to participate. This is connected with the methodological challenge because methodological requirements are limited, in part, by the available technology.

3. **Ethical** (Section 2.5). How to collect data so as to preserve privacy of the users while following accepted experimental practices. How to share the data to allow other researchers to reproduce build upon the results, while maintaining ethical requirements. This is connected with both methodological and technical challenges because ethical requirements have to be satisfied at all times.

We offer a partial answers to the challenges presented above, and present the solutions employed in our previous research.

Instructions on how to acquire code and data needed to reproduce analysis from this thesis are in Appendix A.

2.1 Related work on data collection on Facebook

Facebook is still the most popular online social network with over 2 billion users as of 2018, and provides a valuable resource for investigating social phenomena. Since its origin in 2005, there were three main approaches for collecting Facebook data were. At first, a complete retrieval of regional Facebook data:

- Lewis et al. [3] collected 1640 user profiles, almost all freshmen students of one private college in the Northeast U.S.
- Wilson et al. [57] collected around 10 million users from 22 regional networks, including London, Australia, Turkey, France, Sweden, New York ...
- Viswanath et al. [58] collected 63731 user profiles from New Orleans regional network

Second approach is to access Facebook's internal database, which is usually only available to internal researchers:

- Eckles et al. [59] performed a three week experiment which involved around 48.9 million Facebook users
- Kramer et al. [2] performed a one week experiment involving 689003 Facebook users as participants
- Ugander et al. [56] collected data on over 721 million Facebook users over the period of 28 days

Third approach involves collecting data through an external application using Facebook Graph API:

- McAuley and Leskovec [60] collected Facebook ego networks with the total of 4039 users
- Aral and Walker [61] collected data on over 1.3 million Facebook users through the

Facebook application

- Kosinski et al. [62] collected data on 58466 users from the United States, obtained through the myPersonality Facebook application
- Bohn et al. [63] had a potential to collect data on over 1.3 million users of their Facebook application, but due to technical difficulties were only able to retrieve data on 1712 users
- Jalali et al. [64] collected data on 4683 Facebook users that signed an online petition over a period of 71 days

The first approach was only possible up until 2010, and the second is not available to researchers external to Facebook. We advocate the third approach as the most appropriate for the wider research community.

2.2 Collecting online social network data

Our online survey applications were hosted as a separate web pages which used Facebook Graph API [55] for authenticating users. The survey application way hosted on a separate server which run the survey web interface and a database which stored data on users. Having some form of user authentication is crucial for an online survey application because it prevents multiple voting by the same user, and allows to track users in order to enhance their engagement with the application. Facebook Graph API is particularly convenient for this because users can use their existing Facebook credentials in exchange for the data that Facebook provides, most important being Facebook friendship relations between users from which we constructed social network used in our inference methodology.

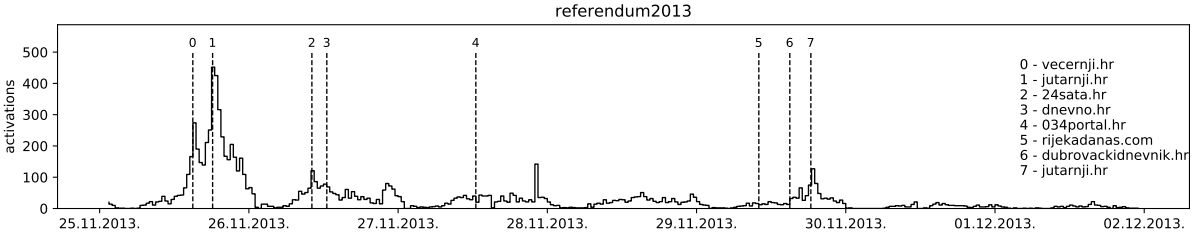
Prior to 2013 Facebook Graph API allowed access to all friendship relations from a registered users, which meant also relations towards users that maybe never heard about the application and had not given informed consent for the usage of their data by third party applications. This was considered a privacy breach and the API was changed in late 2013 to allow access of friendship relations only between registered users. However, for our purposes the friendship relations between registered users was enough as these are potential carriers of the endogenous influence. A year later Facebook also changed its Graph API so that it assigns ID's which are specific for each application, rather than universal for all applications. This makes it much harder to associate users across different Facebook applications, a practice that previously allowed application developers to easily share data on users which is in violation of Facebooks privacy policy [65].

In addition to the friendship relations, the Facebook Graph API also allows retrieval of the basic demographic information such as age and gender. It should be noted that it is still required of users to give an explicit permission for their demographic data to be

collected. The permission is given through Facebook's own authorization dialog provided by the API's interface. However, applications leveraging Facebook Graph API should nevertheless have their own privacy policy and terms of use which they show to their users prior to their registration on the application. We did collect demographic data for our first application referendum2013 but then decided to cease this practice due to potential privacy concerns. Even though users are never identified with their full name in the collected dataset, using only demographic data could lead to potential *deanonymization* - identifying specific users in the dataset. Demographic data also potentially allows the alignment of multiple datasets collected with different applications leveraging Facebook API. This is a kind of indirect deanonymization because it identifies pairs of identical users in otherwise unrelated datasets, although their exact identity is unknown. In theory, existence any information which is shared between two or multiple datasets, demographics being the most notable one, raises the possibility of partial or full deanonymization. This is because it is then possible to cross-identify users in the two datasets based on this shared information, and to gain more information on each user than what is contained in each individual dataset alone. In the end, this could lead to a full deanonymization of some users. In the cases of sabor2015 and sabor2016 applications when we do not collect demographics data the only information shared by both our application and the Facebook itself are the friendship relations between users. Registration times of users and referral links which are logged on our web server are exclusive to our application and by itself cannot aid much in deanonymization.

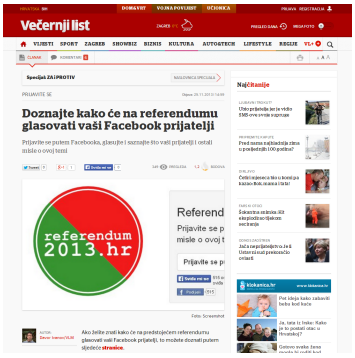
In the end, collecting demographic data through our first survey application - referendum2013, did allow us to perform exploratory analysis of referendum2013 dataset and asses whether or not our collected data is representative of Facebook population. More details on this are available in Section 2.3. Section 2.5 explains in more detail how ethical challenges for Facebook data collections changed in the past couple of years.

Once users register on the application, they can cast their votes on the survey, share link to the application through Facebook, and see summary statistics for their Facebook friends as well as for all other registered users. In order not to influence the vote of the user the application's interface displays summary statistics only after they actually voted. As an incentive for users to share the link to the application through Facebook, in sabor2015 and sabor2016 applications we also displayed a number of their Facebook friends that visited the application by following the link on their share. This number was compared to other users and the application reported a rank among all users. As this rank as well as summary statistics changed throughout the period during which application was active, the users had to repeatedly return to the application which prolonged user's engagement. To preserve privacy of their Facebook friends the summary statistics for their friends

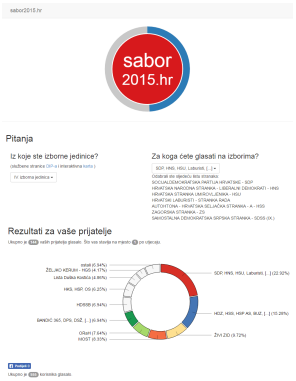


(a)

(b)



(c)



(d)

Figure 2.1: News media coverage for our online survey application and a screenshot of a survey web interface. An example of two online news media articles that reported on our referendum2013 survey application - an article from news portal jutarnji.hr * (Panel 2.1b) and an article from news portal vecernji.hr † (Panel 2.1c). Panel 2.1a shows the number of users registered on referendum2013 survey application in 30 minute intervals, annotated with major online news media articles which reported on the application. The first two peaks (numbered 0 and 1) correspond to the articles in Panels 2.1b and 2.1c. The window of 30 minutes is also used as a window size in the inference methodology. Some of the peaks in user registrations correspond to the publication of news articles which indicate possible exogenous influence. Panel 2.1d shows the screenshot of the web interface for the sabor2015.hr online survey application, which is similar to the ones used in both referendum2013 and sabor2016 applications. All three applications allowed registered users to cast a vote for the upcoming referendum or elections, share the link to the application through Facebook, and see summary voting statistics for their Facebook friends and all registered users.

was not displayed unless more than certain number of their friends voted on the survey. Survey applications were active a week or two before the actual pooling day (Table 2.1) which aided us in attracting new users as the surveys provided a way for the news media and general public to asses the possible election outcomes.

In order to more systematically track the media coverage for our referendum2013 application we also manually collected data from Google Analytics which contained number of users visiting our survey application through an external websites such as online news portals (Figure 2.1a). For sabor2015 and sabor2016 applications we did not have to collect data in this way because we collected referral links from users directly, which essentially contains the same information as Google Analytics but on a much finer scale and for each user individually instead of aggregate estimates. Knowing how many users visited our survey application from a specific external website gives us an opportunity to estimate exogenous influence these media sources had on the number of registrations.

2.3 Exploratory analysis of the collected social network data

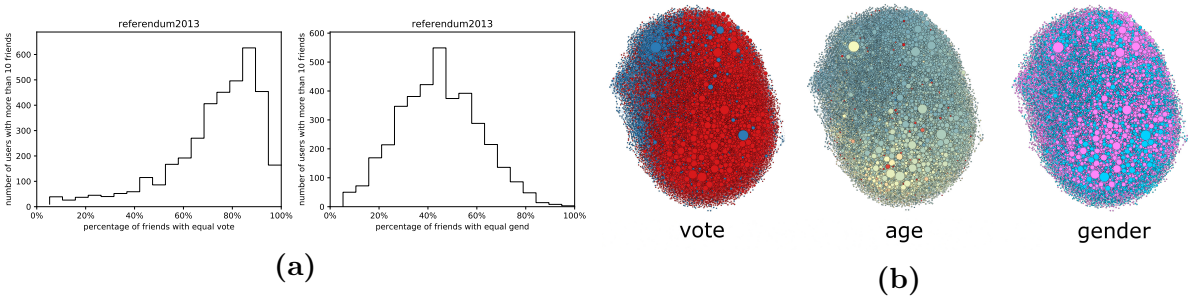
In the following section we show descriptive analysis of the three Facebook datasets that we collected. Datasets consists of the Facebook friendship connections between users that registered on our online survey applications, exact times of their registration, and for some datasets - demographics data and referral links from which users visited our application (Table 2.1). Of course, we also have survey responses for each user that responded to the survey. We only collected demographics data during our first survey related to referendum in 2013 - the referendum2013 survey. For the subsequent surveys related to the parliamentary elections in Croatia in 2015 and 2016 - sabor2015 and sabor2016, we decided to rather collect data on referral links. These were much more useful for us because they effectively give us origin of users - whether they visited our application by visiting a link from Facebook or some other external website. We will later use this as a gold standard data to evaluate our inference methodology. Demographic and survey response data could be used to build more complex models of influence by correcting for the potential confounder variables, for example gender, age, and similar political preferences which could conflate influence with correlation effects in data. As our model of influence (described in chapter 3) incorporates only registration times of users and their mutual friendship connections, and due to the privacy concerns (Section 2.5) we decided not to collect demographics data after our first survey.

As the first online survey that we did - the referendum2013 survey, is the only one for which we collected demographics data, we decided to perform basic exploratory analysis

Dataset	Time period	Users	Collected data	URL of application
referendum2013.hr	25.11. - 1.12.2013.	10175	friendships, demographics	https://github.com/devArena/referendum2013.hr
sabor2015.hr	2.11. - 8.11.2015.	6909	friendships, referral links	https://github.com/matijapiskorec/sabor2015.hr https://bitbucket.org/marin/sabor2015.hr
sabor2016.hr	5.9. - 11.9.2016.	3818	friendships, referral links	

Table 2.1: Summary statistics of the collected social network datasets. Surveys were active typically one week prior to the actual pooling day, and the exact period is indicated in the column "Time period". Friendship connections and demographic data were collected using Facebook Graph API, following user's explicit permission after they authorized with their Facebook credentials. Referral links were collected using our own web server that was hosting the survey application. They indicate user's origin, information which we use to evaluate our inference methodology. Source codes of the referendum2013 and sabor2015 applications are freely available on Github open source code repository, and the corresponding links are indicated in the "URL" column. Table reproduced from [66].

(Figure 2.3). By plotting the distribution of the number of friends that share user's survey response we can estimate an amount of *political homophily* in the Facebook friendship network (Figure 2.2a). We can see that majority of users have 80% or more Facebook friends that voted the same as they did, which indicates that friendship networks are very homogeneous with respect to the political orientation - we tend to associate with users that share our political views. Whether this is a purely *correlational* effect arising from chance is another question, but this fact indicates presence of potential contributing factor for endogenous influence between users. Also, by observing the actual communities of political orientation in friendship network (Figures 2.4 and 2.6) we can see a large *polarizing* effect - users are clustered in two distinct communities based on their political orientation. Running a Louvain multilevel algorithm for community finding [67, 68] identifies 27 communities in the referendum2013 network. They are all highly homogeneous with respect to votes - almost all users in each particular community have an identical survey response. Also, their registration dynamics is very similar and resembles the global registration pattern. Two of such communities are shown in Figure 2.6. The same Figure also shows a community whose registration pattern is very different than the global registration pattern, with a distinct peak in user activity at one particular hour approximately at the middle of survey period. This community is also highly heterogeneous with respect to votes, having approximately equal number of users of both political orientations. This peak of activity is not present in other communities and does not follow after a publication of an online news article, which suggests it is probably driven purely by the endogenous influence. The homophily with respect to age is also pronounced, with users being more likely to friend other users that are close to them in age. Similar effect is also observed in a much larger sample of Facebook users [56]. On the other hand, homophily with respect to gender is almost nonexistent. Users are equally likely to friend other users of both gender. An example of data from sabor2015 dataset which stores information on user



(a) (b)

Figure 2.2: Homophily of the Facebook friendship network in the referendum2013 dataset is shown on Panel 2.2a, which shows the distributions of the percentage of user's friends that voted the same (left) or are of the same gender (right). Users are more likely to friend other uses that voted the same as them (in this case we only have two possible votes "for" and "against") - the distribution (left) is shifted to values higher than 50%. On the other hand, users are equally likely to friend users of both gender - the distribution (right) is centered around 50%. We can bring similar conclusions by watching Facebook social network visualizations on Panel 2.2b which are colored by vote (blue for "for" and red for "against"), age (pale blue for for voters bellow 30 years of age, pale yellow for middle age voters and orange-red for voters above 50 years of age), and gender (pink for female voters and blue for male voters). Homophily with respect to age is shown on Figure 2.3 which plots the age distribution of friends separately for several age groups of users. Size of the nodes correspond to the number of Facebook friends each user has in these networks. For attributes where there is high homophily - votes and age, we observe clustering of users into compact communities based on these attributes.

sessions is shown on Table 2.2.

User id	Time login*	Time share	Referrer id**	Referrer class**	Friend count	Election list id
0	4798	-1	-1	facebook	363	37
1	5684	5691	-1	facebook	88	8
2	2099	-1	3145	facebook	485	37
3	4073	-1	4816	facebook	861	4
4	5471	-1	-1	facebook	108	8
5	1106	-1	-1	facebook	53	4

* used in inference, ** used in evaluation

Table 2.2: An example data from sabor2015 dataset which shows information on user sessions. Each user session corresponds to the first registration of a specific user, with information containing the times of login and first share (columns "Time login" and "Time share", in minutes since reference time), origin of the user (column "Referrer class", obtained from referral link), whether or not it came through a Facebook share of another user and which user it was (column "Referrer id") and survey response (column "Response id", in this case a vote for one of the political parties). Friendship network dataset is available in GML format and as an edge list. Table partially reproduced from [66]

2.4 Methodological challenges

Online surveys built on top of the popular online social networks give us an opportunity to perform large scale polling with relatively low budget, in any case much larger than would be possible with traditional survey methods. However, there are methodological challenges which we summarize in three main points: 1) representativeness, 2) engagement, and 3)

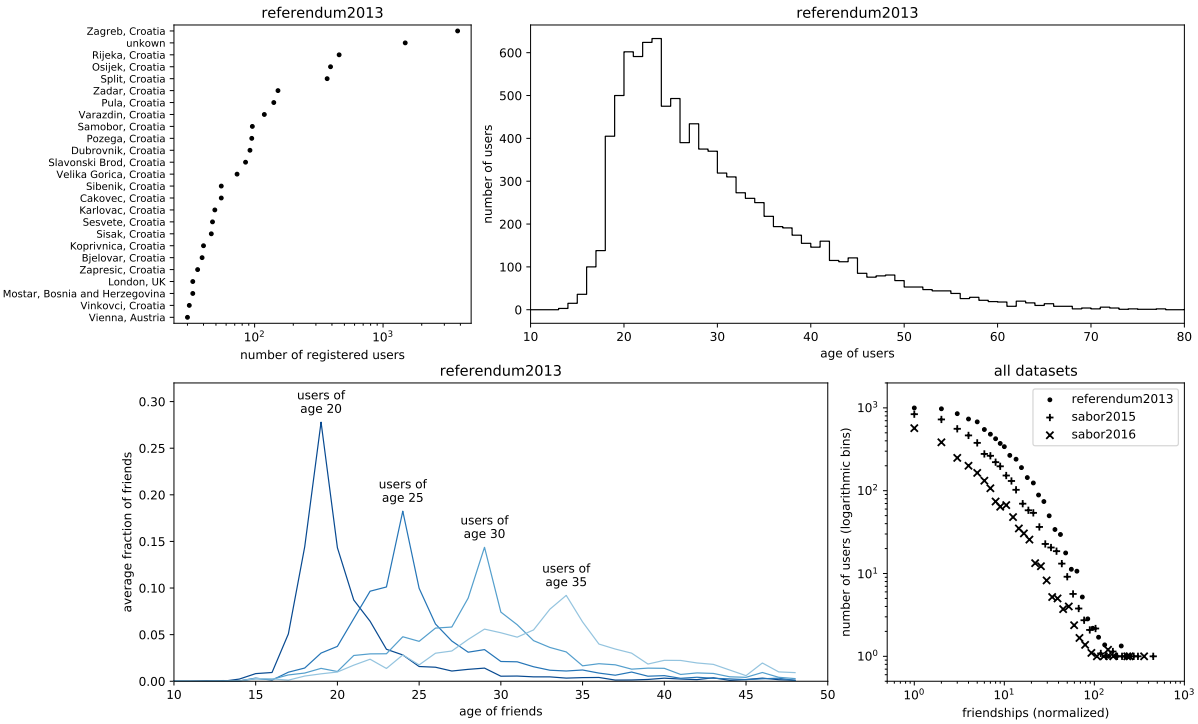


Figure 2.3: Exploratory analysis of the referendum2013 dataset reveals several characteristics. Self-reported locality information (top left) shows that majority of users come from Zagreb, Croatia. This is expected as we did not restrict participation on the survey based on user's location. This allowed us to also obtain responses from Croatian citizens living abroad. We believe that the language of the survey (Croatian) itself served as the most effective filter for our population of interest. Age distribution (top right) show that the majority of registered users are between 20 and 30 years of age, which is much younger than what could be expected from the general population. Panel on bottom left shows age distribution of friends separately for several groups of users of different age. It shows how users are much more likely to friend other users that are of similar age as them. This homophily with respect to age was already shown in social network visualization in Figure 2.2b. The degree distribution of social networks from all three datasets (bottom right) show a scale-free property - majority of users have a relatively low number of friends while there are couple of highly connected users [‡]. The fact that some of these statistics deviate from the ones we would expect on the general population does not influence our methodology much as we are more interested in obtaining a representative sample of Facebook population rather than general population, and all of these statistics are in accordance with the ones obtained from the whole Facebook network [56].

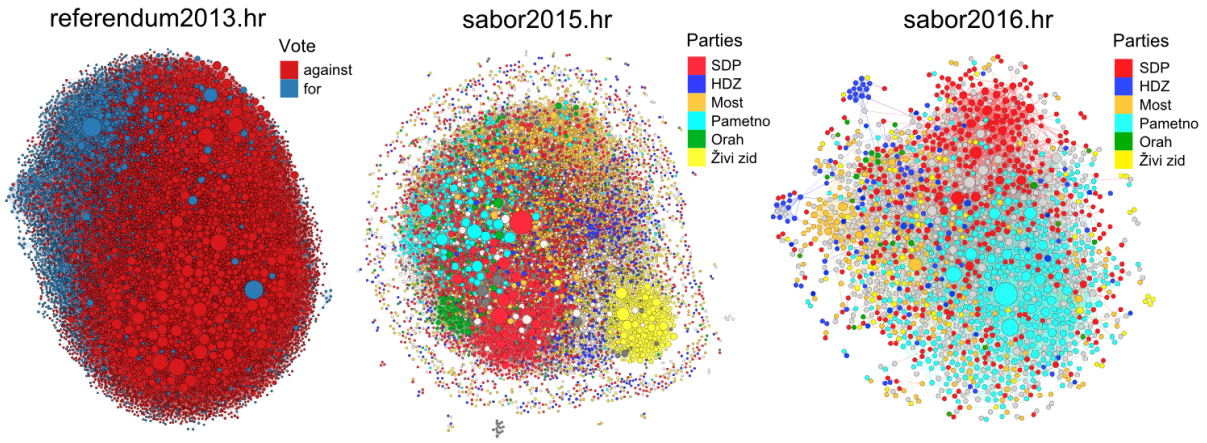


Figure 2.4: Collected Facebook friendship networks of users who registered on three of our Facebook online survey applications: referendum2013.hr (10175 registered users), sabor2015.hr (6909 registered users) and sabor2016.hr (3818 registered users). Nodes are colored based on the survey response and sizes correspond to the number of their Facebook friends that also registered on the same application. Similar as in Figure 2.2, clustering of users into communities based on votes shows a homophily effect - users that share political preferences (which are, to a degree, reflected in their survey response) are more likely to form Facebook friendship connections. Whether the act of forming a friendship connection came as a *cause* or a *consequence* of their political preference or any other characteristic (for example, see Figure 2.2 for age and gender) is another question altogether. Regardless, this might suggest a potential endogenous influence which we try to elucidate with our inference methodology.

causation.

The first question is the one of *representativeness* - Are we collecting a representative sample of our population of interest? In polling literature, spreading the survey organically through the personal (online or offline) connections of respondents is called *snowball sampling* [69]. In Section 2.3 we showed that our collected network of users has representative aggregate properties - degree distribution, gender distribution and age distribution of friends are qualitatively similar to the ones obtained from the whole Facebook network [56]. This suggests that we might have a representative sample of Facebook users, although not necessarily a representative sample of a an underlying population. Regardless, an issue of representativeness is not crucial for the estimation of influence.

The second question is the one of *engagement* - How to engage users to respond to a survey? Again, sharing information with your peers over Facebook is closest analogue to snowball sampling [69] in traditional poling, with the power and reach catalyzed by digital technology. However, reaching large number of respondents is still not guarantied. Crucial factors are novelty and engagement - how are you motivating your participants to share the application to their peers? For example, myPersonality project [62] started with a seed of around 150 users but over the course of four years it spread to engage over six million other users. Subsequent applications, some of which featured better design and features, failed to attract so many participants. The third question is the one of *causation* - Can we

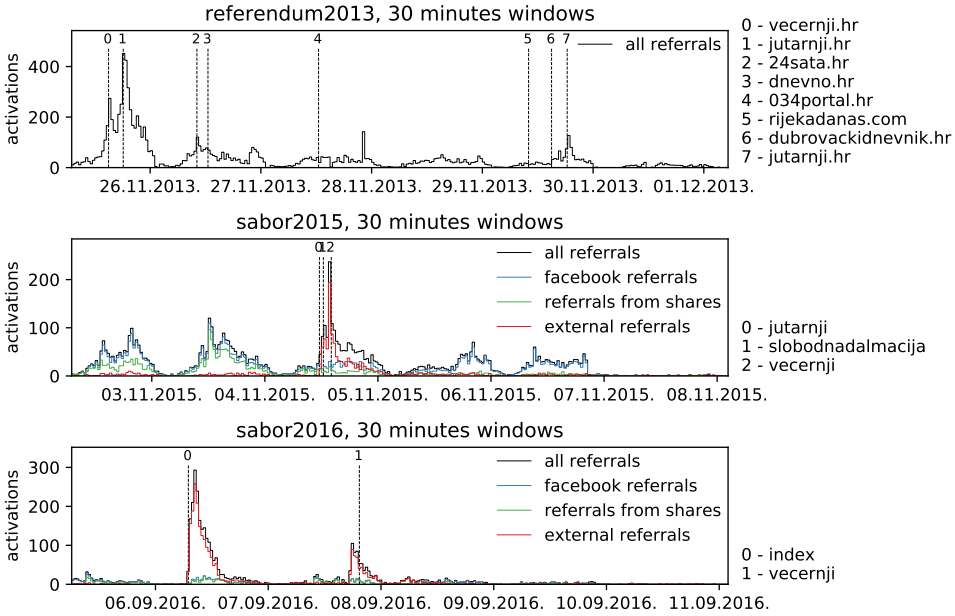


Figure 2.5: Collected registration times of users who registered on three of our Facebook online survey applications: referendum2013.hr (10175 registered users), sabor2015.hr (6909 registered users) and sabor2016.hr (3818 registered users). Registrations are binned into a 30 minute time periods - the same what we use in our inference methodology to determine active and inactive users. Histograms are annotated with coverage of our survey applications by major online news providers. Some of the news coverage align closely with the sudden peaks in user registration which indicates possible exogenous influence. For sabor2015 and sabor2016 we also collected referral links from which users visited our survey application (from which we derive "Referrer id" and "Referrer class" variables in Table 2.2). We separate these based on whether they originated from Facebook or an external website. Those that originated from Facebook could additionally be associated with a share from another user, which is a strong indication of endogenous influence. Those that originate from an external website are a strong indication of exogenous influence. We use this information as a form of gold standard labels in evaluation of our inference methodology in Chapter 5.

measure causal effects on purely observational data? While Facebooks itself offers troves of observational data to researchers able to access its internal database, to measure causal effects it is needed to perform an experiment. And conducting an experiment in digital domain faces the bottleneck of informed consent, a problem which we deal in more detail in Section 2.5. Experiments could be performed using Facebook Graph API which allows external researchers to obtain limited data with the consent of the user. Researchers have full freedom to design their experiments an choose which variables to record and store.

The large volume of data we can collect from online social networks can help us reduce variance, but it does not influence bias at all! If our sampling method is biased we will just be more precise in our wrong measurement. In such complex social datasets it is very unlikely that we will find strong, simple-to-explain, universal effects. It is more likely we will want to concentrate on a specific group of users to maximize our predictive power, which practically means lowering our sample size significantly until we are dealing with

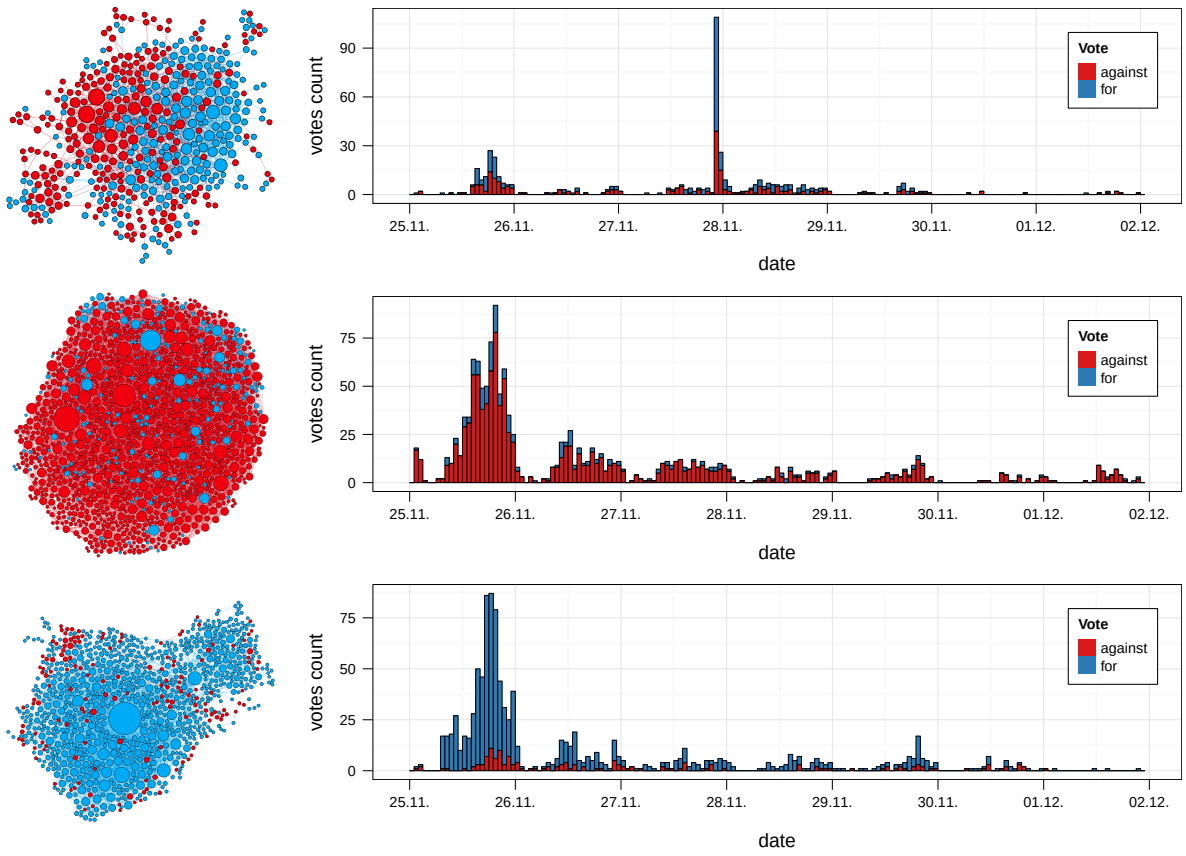


Figure 2.6: Three examples of friendship communities (out of 27 which we identified) between users of our online survey application referendum2013, obtained with multilevel algorithm for community finding [67]. Nodes are colored based on user's responses - red for "against" and blue for "for" votes, and the size of the nodes correspond to the number of friends each user has. Panels on the right show number of registered users in each hour throughout the data collection period. The middle and bottom communities feature high homogeneity of the survey responses which is typical for the majority of communities that we identified. Each of them also has a couple of highly connected users that might serve as the drivers of the endogenous influence. Top community is an interesting exception because it features almost equal number of users of both political orientations, and has no highly connected users. The registration pattern is also atypical - most of the registrations happen during one distinct hour , Community in the top panel is an interesting exception because it has almost equal number of votes for each side, and has no highly connected users. This community also exhibits interesting voting dynamic because majority of its users voted during one particular hour on the evening of 27^{th} of November. Our analysis shows that this peak in activity is characteristic only for this community, which makes it highly likely that it originated because of the peer-driven influence inside this community.

the average-size dataset.

A particular challenge which we faced while collecting user data through our Facebook application is the issue of *observer bias*. Due to the restrictions of the Facebook Graph API we were only allowed to collect friendship relations between users which *both* registered on our application eventually, i.e. until the end of the collection period. This makes our collected friendship network much smaller than the true underlying Facebook social network. What is more, the set of inactive users actually shrinks as we approach the end of observation period. As our assumption regarding exogenous influence is that it acts uniformly on all yet inactivated users, because their number artificially shrinks in our dataset, the magnitude of exogenous influence increases. This is because all of the users in our observed subnetwork eventually register (otherwise we would not collect their data) which does not happen in the real Facebook network. We correct for this observer bias with an additional term in our inference method (Section 4.4).

2.5 Ethical challenges

Although ethical standards in digital social research are still not well defined, we tried to follow recommended guidelines during design and execution of our Facebook survey [14, 15]. Developers of Facebook applications are required to follow Facebook's Platform Policy [§] which defines the conditions under which data could be collected, as well as requirements and responsibilities with regard data's usage and further dissemination to the third parties. For example, selling user's data to third parties is strictly forbidden, even if users (knowingly or not) gave a permission for it. Developers are also required to display terms of use and privacy policy to users *before* they register with their Facebook accounts, as the collection of user data happens automatically upon the registration. In addition, our application also displayed a short version of the terms of use and privacy policy on the front page of the application which was visible to nonregistered users. Our privacy policy stated that we will use data strictly for research purposes and that we will provide anonymized version of the data to the research community. Original versions of the terms of conditions and privacy policy, in Croatian, are available on the Github open code repositories on which our first two survey applications - referendum2013 and sabor2015, are stored. Links to the repositories are in Table 2.1, and the full texts of the policies are also reproduced in Appendix B.

Since the time of our first data collection in December 2013, the general public and research community is more aware of the potential for misuse of data collected from online social networks. Probably the most contributing factors to this are major publicized

[§] `https://developers.facebook.com/policy`

scandals related to U.S. government surveillance as exposed by Edward Snowden in mid 2013 [70], as well as Facebook privacy breach scandal related to the Cambridge Analytica company in early 2018 [71, 72, 73, 74]. In the academic community, notable examples of violating ethical guidelines during research conducted on Facebook data include "Tastes, Ties, and Time" study [3] and the study of emotional contagion [2]. In the former study the data was downloaded freely from Facebook and in the later data was accessed internally from Facebook. In both cases participants did not give an informed consent and were not given an option to opt out from the study. The reason for that was logistic rather than purposeful misconduct - in the first study Facebook regional data was scrapped directly from Facebook without participants knowledge, and in the second study the scale of the study was so large (millions of participants) and an act of informed consent was likely to compromise research goals. For more info on the reactions to these studies see [75] (emotional contagion) and [76] (tastes, ties, and time).

Online book "Bit By Bit: Social Research in the Digital Age" [14] gives very specific guidelines for conducting ethical digital social research, including:

- Respect for Persons is about treating people as autonomous and honoring their wishes.
- Beneficence is about understanding and improving the risk/benefit profile of your study, and then deciding if it strikes the right balance.
- Justice is about ensuring that the risks and benefits of research are distributed fairly.

In our case, the informed consent for the participants in our online survey application was elicited on two levels. First, the front web page of our survey application, next to the registration button, featured a disclaimer that informed users which data will be collected by the survey and how it will be used. Second, once user chooses to register on our application with his Facebook credentials he is redirected to Facebook's own interface dialog which informs him which Facebook data will our application collect, and presents him with links to both Facebook's Platform Policy and the privacy policy of our application. Also, users are able to opt out from delivering their data. This second step is managed by Facebook API interface and is a standard procedure for all third-party Facebook applications. In addition to these, there are also separate web pages, accessible to both registered and unregistered users, with information on the Frequently Asked Questions (FAQ) and the terms of use regarding our survey application. This two-step process is necessary because Facebook API interface manages explicitly only Facebook-derived data. In our case these are Facebook friendship relations and, in the case of referendum2013.hr application, demographic data such as age and gender. In addition to these, we also collect user registration times of users and referral links from which they visited our online application. Collection of these is explained in the disclaimer and our privacy policy to which official

Facebook API interface links. The full texts of the disclaimer and the privacy policy are available in Appendix B.

We decided to collect only data which was absolutely needed for the research, and when we planned to release our data we decided to provide all safeguards as to protect the privacy of the users, even if this meant releasing less data than what we initially collected. Even if we remove personally identifiable information [77] from our released dataset it does not mean we have effectively anonymized the data, as partial deanonymization is still possible, sometimes by aligning the data with some external source which is not completely anonymous. Regarding the benefit to our users, they were given summary information on their friends which they would otherwise not be able to obtain. The society as a whole was given the results of the global survey. Regarding data dissemination, we decided to follow established practice [78] and only share our data with individuals that signed an explicit Data Access Agreement. The purpose of the agreement is to satisfy requirement for reproducible research while, at the same time, protecting the privacy of users and respecting the policies of the online social network service provider. The full text of the agreement is available in Appendix B.

Although immensely powerful in terms of data collection potential, online social networks still provide a challenge for experimental design [15, 79]. The bottleneck is the informed consent - requiring users to give an explicit consent for participating in the study or, at least, giving them an opportunity to opt-out of it afterwards. This significantly reduces amount of data researchers can collect, as users are increasingly aware of threats to which they are in turn exposed. Even when researchers are in a position to automatically present their study to the large fraction of Facebook users the number of responses is usually just a fraction of an initial reach. For example, a study from Aral and Walker [61] presented its study on a sample of 1.3 million Facebook users but still managed to receive responses of only 7730 users. An alternative is doing an observational study.

Part of the problem from the legislative side is that the online service providers are not research institutions and as such are not obliged to follow standard experimental practices, for example the “Common Rule” which states that participants should always be granted an opportunity to opt-out from the experiment. If such provisions are not provided, scientific publishing of such data might be problematic. Still, future promises some interesting developments. Providers of the most popular online social network services are trying to consolidate requirements of the industry with the established academic practices [80]. This is important because institutional review boards (IRBs) are tailored for academic institutions and traditionally research oriented companies (for example, pharmaceuticals), but not for the new data-oriented companies. The introduction of General

Data Protection Regulation (GDPR) by the European Union, which aims to give users more control over their own data and increase transparency in terms of data handling by the data providers, forced major online social network providers to change their data handling policies globally. Also, a recent Data Transfer Project [¶] initiative from leading online social network providers - Facebook, Google, Twitter and Microsoft, should allow seamless transfer of user data between their platforms.

[¶] https://datatransferproject.dev/

Chapter 3

Models

In this chapter a brief overview of the related work regarding modeling influence in social networks is given. Modeling influence is closely related to the modeling of *contagion*, *spreading* or *diffusion* processes in networks which arise due to the interactions between individuals. In the remainder of the chapter the two approaches are presented for modeling endogenous and exogenous influence in online social networks in case when only available information is the social network between users and a single activation cascade. First approach is a direct one where endogenous influence is modeled with an exponential decaying function, exogenous influence is modeled indirectly and the inference of parameters is done manually. Second approach is methodologically more principled - both endogenous and exogenous influences are modeled with explicit microscopic influence models and the inference is performed through maximum likelihood method. For this a log-likelihood function is used which gives the probability of observing particular activation cascade in the online social network as a function of model's parameters. Maximum likelihood inference method is explained in detail in chapter 4.

3.1 Related work on modeling influence

Some of the most commonly used models for information diffusion are inspired by epidemiological models which model how a disease spreads in a population [81, 82, 83]. These models of biological contagion are often *compartmental models* [84] where nodes can take one of several states of *compartments*, and there are conditions under which nodes change their state from one to another. Three most typical biological compartmental models are: *susceptible-infected* (SI), *susceptible-infected-susceptible* (SIS) and *susceptible-infected-recovered* (SIR). Each of these *microscopic* models defines corresponding compartments and transition rules which are dependent on various factors - for example, number of already infected peers. Such is the case with the transition from the susceptible to

infection state which is more probable is more of the peers are themselves infected. Transition rule can also be independent on the neighborhood of a node, for example when the transition is *spontaneous* because node transitions from infected to recovered state. In this particular example the "recovered" state is also an *absorbing* state because there is no transition rule from this state to any other. These three models are often used because of their simplicity which makes them analytically tractable. However, in order to gain insight into real epidemics one often has to use more complicated compartmental models which are designed to be as realistic as possible and where parameters are estimated from real data [85, 86].

Information diffusion and social contagion in general could, in some cases, be modeled with biological contagions [83], although recent experimental evidence suggests that there are several crucial differences [41]. Social contagions usually exhibit more complex functional dependencies with regard to the current state of the neighborhood of a node [28, 87]. For example, in classic biological contagion it is common to assume that there is a simple monotone dependency between the number of infected peers and a probability of infection, while in social contagions there is usually a threshold of minimal number of peers needed to transmit a social contagion. Also, the parameter that drives the contagion might be related to the number of connected components in the person's immediate neighborhood, instead of the number of neighbors [87].

Another crucial difference is the treatment of exogenous effects and various forms of social reinforcement. Epidemic spreading is possible to model with simpler contagion models where endogenous factors play a dominant role - for example, susceptibility of an individual to a certain disease and the pairwise transmission rate. This means that the probability of contagion is independent of the neighborhood structure and the state of users in it, as well as any other external factors. However, social contagions often include more complex mechanisms of transmission due to the common presence of various forms of social reinforcement such as reciprocity [88], social feedback [59], and homophily [31] (a tendency of similar nodes to form connections between each other). Social contagion also exhibit presence of *exogenous* factors [49] which could act as a significant drivers of influence, for example political unrest [1, 52], natural disasters [34] and external media [33]. All of these have a potential for confounding with the true social influence [30, 31, 89]. One strategy of decoupling these correlational effects from influence is by using randomization strategies on networks [32], which should diminish true influence and leave correlation intact.

Usage of latent states which are inherently unobservable in data is problematic for inference. This is why it is often more appropriate to use models where all states are observarble - for example, Independent Cascade (IC) model [90] and Linear Threshold

(LT) model [19, 24]. They feature only two observable state - *active* and *inactive* which simply determine whether spreading process already reached a user or not. The simplicity of these two models allows them to be studied analytically [91], and aids in statistical inference from data [91]. They can also be used as a building blocks for more complex applications such as influence maximization [92].

However, over the course of several decades the study of social contagions yielded many different social contagion models [93] and inference of dynamics in social networks [21, 94]. One of the first was Granovetter's threshold model [19], also known as linear threshold model, where a node is activated if the sum of influences from its peers exceed its own influence threshold. In Watts threshold model [20] a node is activated if the fraction of its activated peers exceeds its threshold, which is drawn from a distribution. Generalized model of contagion [95] introduces the memory of past exposures which influences contagion, and can be used for both biological and social contagion. This model was motivated by the need to more finely distinguish between two extreme cases: (i) where successive contacts result in independent probability of infection, for example like in compartmental models and (ii) where there is a fixed threshold of contacts after which probability of infection immediately changes. Centola-Macy model [28] is similar to Watts model, but uses absolute number of activated peers instead of their fraction. Ignorant-spreader-stifler (ISS) model [96] is similar to SIR compartmental model with a difference that a transmission to absorbing state (stifler) is not spontaneous but depends on the presence of spreaders or stiflers in the neighborhood of the node. In multiparametric model [97] an activation of a node depends on the weighted linear combination of three terms: (i) personal preference, (ii) an average of its neighbors states and (iii) average of all nodes in the network. In multi-stage complex contagions [98] a node can achieve and additional hyperactive state where exerts bonus influence along with the regular influence. In synergistic model [99] infectivity and susceptibility of a node is dependent on the number of active peers. Finally, in voter model [100] at each time step one node is chosen uniformly at random from the network and it adopts uniformly at random a state from one of its peers.

3.2 Modeling exogenous influence directly

First, a simple method of estimation of endogenous and exogenous influence is demonstrated where a single endogenous influence model is defined and which then uses a threshold rule to differentiate between these two types of influence [101]. In a way, this method does rely on explicit inference method but rather estimates magnitude of influences directly from data. Influence between users is modeled through activation probability - each

activated user has a potential to activate, in the next time step, each of its peers with probability p_0 which decays exponentially in time *. Because individual activations are independent of others this makes it a form of IC model. The rate of the exponential decay $p_0 e^{-\lambda t}$ is determined by the decay parameter λ. For each user i at time t, the probability of endogenous activation can be expressed as the probability of being activated from any of its already activated peers $N(i)$:

$$p_i(t) = 1 - \prod_{k \in N(i): t_k < t} (1 - p_0 e^{-\lambda(t - t_k)}), \tag{3.1}$$

Here, t_k is the activation time of peer k that activated before time t. Later, in Section 3.3, we will use the same exponential decay model (Equation 3.3) in the joint inference method. The assumption is that user activations are due to superposition of endogenous and exogenous influence, and that there is a statistical difference between these two influences which can be observed in data. In Section 4.2 it will be shown that on simulated activation cascades a simple threshold rule could be used to differentiate between these two influences, based on differences in values of $p_i(t)$.

Section 5.1 contains the results of evaluation on the empirical data obtained from referendum2013 online survey. It also shows how to infer the influence decay parameter λ. However, the limitation of this model is that the endogenous and exogenous influence are indirectly coupled, and only the endogenous influence is modeled directly. So in the next section a fully probabilistic model of influence is devised which jointly models both the endogenous and exogenous influence.

3.3 Modeling exogenous and endogenous influence jointly

In this section a joint model for endogenous and exogenous influence is presneted which rests on two assumptions (Figure 4.3): (i) endogenous influence depends on the friendship network structure and which users are already active or not, and (ii) exogenous influence is independent on the friendship network structure and the state of users in it, and is constant across all users. Additional assumption is that the parameters of endogenous influence are constant throughout time, while the parameters of exogenous influence may vary. The reason behind this that, although individual probabilities of endogenous activation change for each user depending on the state of the friendship network and the progress of the

*In this formulation the time is a *discrete* rather than *continuous* variable, which should define all probabilities as *masses* rather than *densities*. Although the former are usually written with capital letters here we choose to write probabilities and time as p and t rather than P and T.

activation cascade, the underlying endogenous activation mechanism is still universal for all users and constant in time. This also results in a single set of endogenous influence parameters which eases interpretation of the model. Exogenous influence could have been modeled with an appropriate time-varying parametric model, but here it is modeled non-parametrically instead, which means that we have a separate parameter of exogenous influence at each time step [35]. A very simple model for the exogenous influence is used - a simple probability of exogenous activation $p_{ext}^{(i)}(t)$ that acts on all inactive users equally at each specific time step.

The endogenous influence could be modeled with any appropriately defined microscopic influence model. Three models are chosen whose variations are commonly used in information spreading research: (i) Susceptible-infected (SI) model, Exponential decay (EXP) model and (iii) Logistic threshold (LOG) model. With these choices both *simple* and *complex* contagion models are covered. In simple contagion models the contagion happens due to a direct interaction of the two users without any additional factors. In complex contagion models the contagion is a result of the conditions present in the network as well as pairwise interaction between users. First two models - SI and EXP are special cases of IC model where each user has a certain probability of activating its peers in the next time step independently of the rest of the network, which makes them an example of *simple contagion* models. The difference between the two models is that in SI model this probability does not change in time while with EXP model the probability of activation decays in time, which lowers the influence of your peers that activated farther away in time. The assumption of *decaying influence* is commonly included in both endogenous and exogenous influence models [102, 103]. On the other hand, the LOG model is an example of complex contagion where the probability of activation depends on the number of your peers which are already active. This requirement of multiple interaction models the mechanism of *social reinforcement* which is a known driving force for product adoption [49].

First, the SI model is presented where probability of endogenous activation $p_{SI}^{(i)}(t)$ for user i at time interval $[t-\Delta t, t]$ [†] is defined as follows:

$$p_{SI}^{(i)}(t) = 1 - \prod_{j \in N^{(i)} \text{active at } t} (1 - p_0) = 1 - (1 - p_0)^{a_i(t)} \tag{3.2}$$

The main parameter of SI model is p_0 - a probability that a particular peer j from the set of all $i's$ peers $N^{(i)}$ will activate user i in the next time step $[t-\Delta t, t]$. As each activation from each of the peer is independent, we can simplify the expression for $p_{SI}^{(i)}(t)$ by using $a_i(t)$ - the number of activated peers of user i at time t. Assumption of the SI

[†]Again, the time variable t is discrete here although Δt is used to designate a time increment.

model is that the probability of activating one's peers p_0 does not change in time. This assumption is more appropriate in epidemiological setting from where SI model originated than in information propagation setting where we would expect that the influence between users decays in time. We can achieve this by adding a parameter for influence decay λ, which leads us to the EXP model:

$$p_{EXP}^{(i)}(t) = 1 - \prod_{j \in N^{(i)} \text{active at } t} (1 - p_0 e^{-\lambda(t-t_j)}) \tag{3.3}$$

Here, the parameters of endogenous activation are p_0 and λ. The p_0 is equivalent to the corresponding parameter from the SI model and determines the probability of user j activating user i at the time of its own activation $t = t_j$. The decay parameter λ determines the how fast does the influence decays - a half-decay of influence will happen after approximately $\log(2)/\lambda$ units of time, which in this case is the period over which we aggregate the newly activated users $[t - \Delta t, t]$. For example, decay value of $\lambda = 0.1$ means that the influence decays to half of its value in approximately three units of time. Both SI and EXP models are examples of IC models - each individual has an independent probability of activating each of its peers. Similar as the assumption of non-decaying influence in SI model, this assumption is also more appropriate in epidemiological rather than social setting. In social contagion we expect that the influence increases with the number of *exposures* to which user is exposed, possibly in a nonlinear way. One possibility is to use a *threshold* - a number of exposures which have to be exceeded in order for the influence to reach a nonzero status. The definition of an *exposure* is sometimes ambiguous - it could relate to the number of peers through which one was exposed to an information or to the number of exposures themselves, which could come even from a single peer. The number of peers is chosen as a measure of exposure and so the LOG model is defined as follows:

$$p_{LOG}^{(i)}(t) = \frac{1}{1 + e^{-k(a_i(t) - a_0)}} \tag{3.4}$$

The parameters of endogenous influence here are a_0 and k and they define the shape of the logistic threshold function which determines the probability of endogenous activation $p_{LOG}^{(i)}(t)$ of user i depending on the number of active peers (exposures) $a_i(t)$ which is calculated from data. The parameter a_0 is the number of active friends you need for the probability of endogenous activation to reach 0.5. The parameter k determines the slope of the threshold. In the case of $k = 0$ the threshold is hard:

$$p_{LOG}^{(i)}(k = 0;\, t) = \begin{cases} 1 & \text{if } a_i(t) \geq a_0 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

This means that the probability of activation $p^{(i)}_{LOG}(t)$ is 1 in case $a_i(t) \geq a_0$ and 0 otherwise. Higher values of k soften the threshold so that the probability of activation is nonzero even in cases when the number of active friends is below the threshold value a_0. A probability of activation of $p^{(i)}_{LOG}$ for $k > 0$ at exposure of a_0 active friends is exactly 0.5, and probability of 1 is reached only in a limiting case of large number of exposures.

Having microscopic influence models for both endogenous and exogenous influence we can join them in one probabilistic expression by defining a *likelihood function*. The likelihood function $\mathcal{L}$ gives us probability of observing data D, which in this case are a friendship network and user's activation times, at a particular time t conditioned on the chosen models of influence p^{peer} and p^{ext}:

$$\begin{aligned}\mathcal{L}(D;\, p_{peer}, p_{ext}, t) = &\prod_{i \in \text{activated at } [t-\Delta t, t]} (1 - (1 - p^{(i)}_{peer}(t))(1 - p_{ext}(t))) + \\ &c(t) \prod_{i \in \text{inactive at } t} (1 - p^{(i)}_{peer}(t))(1 - p_{ext}(t))\end{aligned} \tag{3.6}$$

The likelihood consists of two terms, first one which quantifies the agreement for the users that *did* and second one for the users that *did not* activate in a given time period $[t - \Delta, t]$. Assumption is that each activation had to be due to either endogenous or exogenous influence and there are no other possible influences. Explicit dependence on time t will be removed in the inference phase because likelihood will be estimated non-parametrically - there will be a separate estimate $\mathcal{L}$ at each time increment Δt.

In principle, nothing prevents us from using a more general form of exogenous influence $p^{(i)}_{ext}(t)$ which is user-dependent, or different endogenous influence parameters for different groups of users, but this would increase the number of parameters and make our inference harder. Maximum likelihood inference is described in detail in chapter 4, but here several optimizations are shown which do not change the model specification or the solution of the inference although they make inference more efficient and feasible. First, because likelihood function in Equation 3.6 involves multiplication of many small probabilities, which is likely to result in numerical overflow, we exchange multiplication for summation by log-transforming the likelihood:

$$\begin{aligned}\log \mathcal{L}(D;\, p_{peer}, p_{ext}, t) = &\sum_{i \in \text{activated at } [t-\Delta t, t]} \log(1 - (1 - p^{(i)}_{peer}(t))(1 - p_{ext}(t))) + \\ &c(t) \sum_{i \in \text{inactive at } t} \log((1 - p^{(i)}_{peer}(t))(1 - p_{ext}(t)))\end{aligned} \tag{3.7}$$

This does not change the value of the maximum likelihood parameters due to the

monotonicity of logarithm. Second, for additional numerical stability we do not actually calculate expressions for endogenous activation probabilities in Equations 3.2 and 3.3, but instead calculate equivalent expression using *sum-log-exp* trick [104]:

$$p_{SI}^{(i)} = 1 - \exp\left[a_i \log(1 - p_{SI})\right] \tag{3.8}$$

$$p_{EXP}^{(i)} = 1 - \exp\left[\sum_{j \in \text{activated at } t} \log(1 - p_{EXP})\right] \tag{3.9}$$

By substituting product for summation, the trick allows us to avoid numerical underflow while calculating the product in Equations 3.2 and 3.3, which can happen for users that have a large number of peers. Because quantities under the logarithm are of the form $1-p$, in practice we use a special log1p [‡] function in Numpy for calculation of $\log(1+x)$ which provides more precision when x is small.

The role of factor $c(t)$ in the second term of Equations 3.6 and 3.7 deserves an explanation. The second term determines the agreement with users that did not activate in the given time period $[t-\Delta, t]$, neither through endogenous nor through exogenous influence. But the question is on which users does the exogenous influence acts exactly? We know that our friendship network contains just a subset of users in the real online social network. This makes it likely that the true number of inactive user on which exogenous influence could act is actually much larger than what we observe. This underestimate of the number of inactive users on which exogenous influence could act could lead to the overestimate of the exogenous influence. We call this effect an *observer bias* because it is a direct consequence of the data collection methodology - the friendship network contains only the users that eventually registered (activated) on the online survey application. As we approach the end of the observation period the number of inactive users we observe drops to zero, although there are still many inactive users in the true social network which we do not observe. We can correct for this observer bias with factor $c(t)$ in Equations 3.6 and 3.7 that increases the contribution of the inactive users in the likelihood:

$$c(t) = 1 + \alpha \frac{N_{all}}{N_{inactive}(t)} \tag{3.10}$$

Here, N_{all} is the number of all users in the social network, and $N_{inactive}(t)$ is the number of all yet users inactive users at time t. Correction for the observer bias is explained in more detail in Section 4.4.

[‡] `https://docs.scipy.org/doc/numpy/reference/generated/numpy.log1p.html`

Chapter 4

Inference

This chapter presents an inference methodology used to infer parameters of the models for endogenous and exogenous influence described in Chapter 3. Section 4.2 presents a simple direct inference method that effectively infers only the endogenous influence while any deviation from the expected endogenous influence is interpreted as exogenous, and where parameters of influence are determined manually. Section 4.3 presents a more principled inference method that uses a full likelihood function which includes both endogenous and exogenous influence explicitly, and which can be optimized numerically in order to infer maximum likelihood parameters. Sections 4.4 and 4.6 deal with two important technical aspects of the inference methodology - correction for the observer bias which arises from the way the data on activation cascades is collected, and the scalability analysis which shows how the inference method scales to social networks with large number of users. In Section 4.5 the results of extensive experiments where inference is performed on simulated activation cascades are presented. Section 4.7 shows how the inference methodology can be used to estimate a measure of individual and collective influence, and compare it with structural measures of influence in Section 4.8.

4.1 Related work on inference in networks

Whether we are considering inference of network *structure* or *processes* on networks, we are essentially dealing with models that cannot be easily analyzed with standard statistical techniques. In order to perform statistical inference of networks directly there are several problems which have to be accounted for [105, 106]:

First problem inherent in inference on networks is *granularity of observations*. A realization of a network structure generated by a model is considered as a single observation instead of a set of independent, identically distributed (iid) observations. In terms of structure, the whole network is a single realization of our network generating model. This

prevents us in using standard data partitioning techniques, for example for partitioning our network into training and validation sets in order to validate our model. In terms of processes on network such as information diffusion, in principle each individual information transfer could be considered as an almost independent event. Here, dependence is at best locally restrained to the immediate neighborhood of the person. However, existence of exogenous factors invalidates this assumption, as dependence then extends to the arbitrary large percentage of network. Using a likelihood based approach helps because likelihood measures the agreement of the model with the *entire* observed data, and it allows the evaluation of model fitness and model complexity without the use of independent test set.

The second problem relates to the *node correspondence*. Inference of network structure should take into consideration that a particular labeling of nodes should not change the likelihood of being generated by a particular network generating model. In other words, it should treat *isomorphic* networks as being equivalent. This is not an easy problem, as graph isomorphism is in NP-intermediate complexity class [107]. So in order to calculate a likelihood we have to consider all $N!$ possible permutations of node labelings, which is computationally infeasible. A pragmatic solution to this problem is to perform an appropriate sampling strategy like Markov Chain Monte Carlo MCMC [105].

The third problem relates to *likelihood estimation*. Even without the node correspondence problem, in order to calculate the likelihood that a particular model generated the observed network structure we still need to evaluate the probability of each of the N^2 possible edges in the observed network. Again, using appropriate sampling strategies like MCMC could help, although a much more common approach is to use *aggregated statistics* as a proxy for evaluating different models. These include degree distribution or clustering coefficient in case of structure [108], or response correlations in case of dynamics [109]. Comparing only the aggregate features reduces the discriminative power of model validation [110], but is often practiced because it requires less computational resources and allows the usage of standard statistical methods for evaluation. In general, using a likelihood-based approach allows comparing models in a probabilistically unified way, rather than comparing them indirectly by using a subset of many possible aggregated features [111].

Despite all of these issues, there are methods that successfully infer both *network structure* and *processes on networks*. Some of these will be reviewed in the remainder of the section. We have to distinguish models which implicitly or explicitly use network dynamics for inference of network structure and models of processes on networks. Network dynamics is a rather broad term and includes both dynamics of structure, for example like link formation, and processes on networks such as information cascades. It is commonly

used in inference of network structure, for example in network growth models, models of community formation, and models of network structure inferred from dynamic data such as information cascades in online social networks. Inference of processes on networks includes, for example, epidemic and birth-death processes, biochemical and regulatory dynamics, human trails on the Web such as Web navigation and sequences of reviews.

We first review some of the related research on the inference of network structure. Historically, the sociological studies on human social networks pioneered this research direction, several decades before the development of modern theory on complex networks structure and dynamics [112]. These studies mostly used classical graph theory and concentrated on investigating the role and influence of individual nodes in network rather than global properties of social networks. The most significant of these early models are *actor models* which are mostly used to model conditions under which nodes change their outgoing connections [94]. Actor models are flexible enough to incorporate many sociologically relevant features such as *transitive triplets*, *reciprocated ties*, *indirect ties* and *persistent reciprocity*, and they can be inferred from empirical data with maximum likelihood methods [21].

More recent research direction is the investigation of evolution and fundamental properties of empirical complex networks. This includes development and inference of *network growth* models that are able to reproduce global and mesoscale structural properties commonly found in real world networks [105, 106, 111, 113]. From a probabilistic perspective, a network growth model is actually a probability distribution on a space of all possible networks. By using a maximum likelihood estimation one obtain most probable growth model given the data on network growth [105]. Due to the large dimensionality of the likelihood one has to use MCMC or some other sampling procedure for estimation. Another strategy for dealing with high dimensional likelihoods is to use less data, which is usually discriminative enough for selecting among several predefined candidate models (as opposed to parameter estimation) [113]. We can also use supervised learning methods that learn from aggregated network features in order to identify a growth model that might generated a particular network structure [114]. Maximum likelihood can also be used to design complex models of network growth that are composed out of simpler microscopic principles [111].

Doing inference on networks is a matter of representation, sometimes it is not necessary to perform inference on an explicit representation of all possible pairwise connections between nodes if more efficient representation exists that captures desired underlying structural characteristics. Two such representations of network structure for which efficient inference methods were developed are *Kronecker graphs* [106] and *block models* [115].

Kronecker graphs [106] are recursive models of networks that are expressive enough to

model real networks and to reproduce most of their properties. They rely on the *kronecker product* of adjacency matrices which is successively applied to the *initiator graphs* in order to generate self-similar network of arbitrary size. The large scale structure of the network such as communities and other network properties are encoded in the initiator graph, and these can be inferred from empirical networks [106]. The Kronecker graphs have a multinomial distribution for in and out degrees of the nodes, which for some choices of initiator graphs behaves like a power-law distribution, and they follow the densification power law.

Another efficient representation of network structure are *block models* [115] which encode communities (blocks) in network and their mutual connections. Formally, a block model that contains k blocks is a $k \times k$ matrix M where each element M_{ij} gives a probability that a node from block k_i is connected to a node from block k_j. Erdös-Reny networks are a special case of block models where there is only one block. Inferring a block model from an empirical network implicitly performs *community detection* as the blocks can be directly interpreted as communities on network. However, the inference itself is usually conditioned on a specific number of blocks, and so one has to use various complexity measures to either select number of blocks beforehand or somehow incorporate block selection in the inference itself. For example, minimum description length (MDL) can be used in block model inference as a complexity measure [116] which considers not only the number of blocks but also their relative sizes. Also, there are efficient Monte Carlo methods for inference of block models that from data which do not require a predefined number of blocks [117]. What is optimized in these methods is entropy rather than log-likelihood, and inference is performed in a hierarchical way where every level serves as a prior information for the lower level [118].

In social networks, an underlying assumption is that interactions between persons somehow reflect the underlying social relationships. This means we could use information on network dynamics to infer social network structure. Examples of such methods are *CoNNIe* [119], *NetRate* [120], *NetInf* [121] and *InfoPath* [122] which all use generative probabilistic models for inferring pairwise transmission probabilities between persons in a social network from information diffusion data. Pairwise transmission probabilities could be interpreted as weights on social network connections, and can be used to infer the most probable activation sequence between persons, as an exact information on who transmitted information to whom is usually not available. All of these methods optimize for a maximum likelihood information cascade, although using a different optimization method. CoNNIe and NetRate use convex programming, NetInf uses submodular function optimization and InfoPath uses stochastic gradients. Only InfoPath is able to provide an online estimate in case when network structure is changing over time.

While there are several possible representations of network structure which could be used for inference, inference of processes on networks unfortunately still lacks a suitable representation which would allow inference of a broad range of dynamical models using an unified probabilistic framework [123]. In case of binary-state dynamics, where each node can occupy one of two states, we can use *infection rate* $F_{k,m}$ and *recovery rate* $R_{k,m}$ functions which depend only on the degree of node and the number of its neighbors, and which can describe many binary-state processes like SI and SIS models, Bass and Kirman models and voter models. These rate functions can be used to derive a *master equation* for describing time evolutions of the fractions of nodes in each of the states [124]. Unfortunately, currently there are no proposed methods for inference of these functions from data.

Another suitable representation for processes on networks is with a general network dynamics equation [22] which is able to represent epidemic processes, biochemical dynamics, birth-death processes, and gene regulatory dynamics [123]. In all of these cases the equation models the way a state of a node changes depending on the states of its neighbors, where neighbors encode some kind of a relationship structure between entities. For example, in epidemic process the state of a node is its probability of infection, in biochemical dynamics a concentration of a reactant, in birth-death processes a population at a specific site and in gene regulatory dynamics an expression level of a gene. Inference is done by expanding the equation into Hahn series * and then approximating the leading term of the series by using *transient response*, which describes a response of a system after perturbation, and the *response matrix*, both of which are aggregated features of dynamics. This method infers only the functional form of the model, not its parameters, which might be useful for model selection [109].

One particular kind of processes on networks are *human trails* - sequence of content like Web pages, multimedia and reviews which users consume in succession. The transition from one content to another for each individual user is governed by transition probabilities which could be described by Markov chain, meaning that the transmission probability is determined only by the most recent content in a trail [125]. Hypothesis pertaining the human behavior could be expressed with such Markov chains - for example, a uniform hypothesis states that a person is equally likely to interact with any given content at any given time, while structural and similarity hypotheses give preference to content which are better connected (for example, through hyperlinks) or more similar to the previous one. Efficient Bayesian inference methods allow selection of the most probable Markov chain hypotheses given empirical data [126, 127].

Probably the most potent research on processes on social networks relates to the pre-

*Hahn series is a generalization of the Taylor's expansion that includes both negative and real powers.

diction of information cascades given past diffusion traces, with or without the explicit social network structure available [5]. Of those that use an explicit social network structure, LT model [19] can be inferred from data using gradient ascent method [25], AsIC and AsLT using a maximum likelihood estimation [26], and T-BaSIC model (Time-Based Asynchronous Independent Cascades) using logistic regression [128]. However, reproducing realistic temporal dynamics is still difficult [5]. Methods that do not use any information on social network structure have to impose additional assumptions in order to perform inference. For example, a SIS model can be fitted to data under assumptions that all nodes have the same probabilities to adopt the information and to become susceptible at the next time step [27]. Linear Influence model relaxes these assumptions, and it allows inference of individual influence functions for each node separately in a non-parametric way by solving a non-negative least squares problem using the Reflective Newton Method [8]. Partial Differential Equation based model can predict topological and temporal dynamics of an information injected in the network by a given node, and its parameters can be estimated using the Cubic Spline Interpolation method [92].

4.2 Statistical estimation of exogenous influence directly

This section provides a description of a method for inference of endogenous and exogenous influence directly from data using an exponentially-decaying endogenous influence model (Equation 3.1, Section 3.2). This model gives us a probability of endogenous activation $p_i(t)$ for each inactive user i at each time step t. The inference is based on assumption that there is a statistical difference between endogenous and exogenous influence in terms of values of $p_i(t)$. Let us first define an expected probability of endogenous activation $\mu(t)$ over all inactive users at time t as:

$$\mu(t) = \frac{1}{N} \sum_{i:t_i \in (t,+\infty)} p_i(t), \tag{4.1}$$

where N is the number of inactive users at time t. Exogenous influence is estimated indirectly, as every activation that cannot be explained as an endogenous activation. The assumption is that the exogenous influence is independent of $p_i(t)$ and that it influences all user in a network uniformly. If there is only exogenous influence acting in the network then the set of newly activated users (those that activated in time window $[t-\Delta, t]$) should be a unbiased uniform sub-sample among all inactive users. If there is endogenous influence present in social network, the set of newly activated users should be biased towards users with high endogenous activation probability $p_i(t)$. The number of users activated due

to endogenous influence in time window $[t-\Delta, t]$ is estimated as a discrepancy between expected probability of endogenous activation $\mu(t)$ and actual activation probabilities summed over all newly activated users:

$$peer(t) = \sum_{i:t_i \in [t-\Delta,t]} \mathbb{1}(p_i(t) - \mu(t)), \tag{4.2}$$

Equation 4.2 counts the number of endogenously activated users by classifying all newly activated users as endogenous if their probability of endogenous activation $p_i(t)$ is higher than expected probability of endogenous activation $\mu(t)$, and exogenous otherwise. An actual threshold rule is implemented with the help of indicator function $\mathbb{1}(x)$ which is equal to 1 is argument is non-negative, otherwise it is zero.

This method is evaluated on simulated activation cascades using an actual social network collected with referendum2013 application. For simulating an activation cascade a variant of IC model is used where each active user has an independent probability of activating any of its inactive peers at each discrete time step. This conforms with the exponentially-decaying probability of endogenous activation in Equation 3.1 that assumes independent activations from peers, each of which exerts influence of the form $p_0 e^{-\lambda_p (t-t')}$ towards the user. The simulation is initialized by activating a single user chosen at random.

Along the endogenous influence, exogenous influence is also introduced which exerts on all inactive users a certain probability of activation in the next time step which is equal for all users, regardless on how many of their peers are already active. Although exogenous influence is equal for all inactive users *at a specific time step*, it may change in time. For this simulation a spiked exponentially-decaying influence of the form $q_0 e^{-\lambda_e (t-t')}$ is used. This form is qualitatively similar to what can be observed in empirical data (see, for example, Figure 2.1a).

Figure 4.1 shows results on simulated activation cascade on referendum2013 social network with endogenous influence parameters: $p_0 = 0.03$, $\lambda_p = 0.02$ and exogenous influence parameters: $q_0 = 0.2$, $\lambda_e = 0.3$ that fires at 5^{th} and 15^{th} step of the simulation. Using a threshold rule implemented in Equation 4.2 it is possible to estimate the total number of users activated due to the endogenous or the exogenous influence. A simple method is used as a baseline that classifies as exogenously activated users all those that at the time of activation had no active peers. This is a rather conservative measure that tends to underestimate the number of exogenously activated users, especially near the end of the simulation where the majority of users in social network are already active, and it becomes increasingly rare to find any users without at least one active peer.

Figure 4.2 shows the distributions of endogenous activation probabilities $p_i(t)$ for all

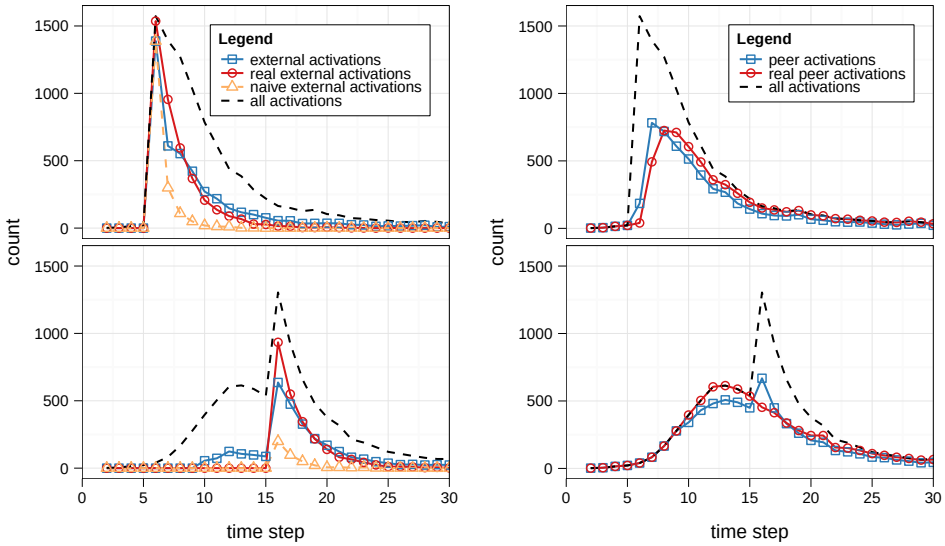


Figure 4.1: Direct statistical estimation of influence on a simulated activation cascade. The underlying social network is a Facebook friendship network of users of referendum2013 online survey application. Activation cascade is simulated using IC model with exponential decay of endogenous influence between users. The shape of the exogenous influence is a single exponentially decaying peak which acts on all inactive users in the network, fired at the 5^{th} (top panels) and 15^{th} (bottom panels) step of the simulation. The method of direct statistical estimation is able to estimate the magnitudes of exogenous (left panels) and endogenous influence (right panels) in both cases.

inactive users at the 15^{th} step of the simulation. It shows that the distribution of $p_i(t)$ for all exogenously activated users is distributed as an uniform unbiased sub-sample of $p_i(t)$ of all newly activated users. A baseline for exogenous influence a simple method [35, 120] is used where users are classified as exogenously activated if, at the time of activation, they had no previously activated peers. An issue with this baseline is that it underestimates the number of exogenously activated users, especially near the end of the activation cascade [35]. As more and more user get activated, it is increasingly unlikely that an user will not have at least one active friend just by chance alone. The usage of $\mu(t)$ tries to remedy this as it tracks the average number of active friends each user should have, and raises the criterion for being classified as exogenously activated. Note that if we set $\lambda = 0$ in the endogenous influence model in Equation 3.1 - a case where the influence of our active peers does not decay in time, we effectively obtain the baseline model for exogenous influence.

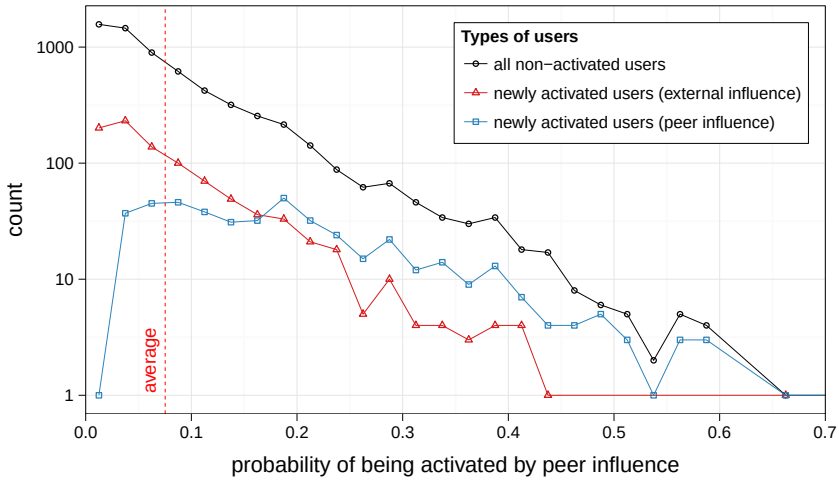


Figure 4.2: Distribution of endogenous activation probability $p_i(t)$ (Equation 3.1) for all newly activated users in simulated activation cascade. The distribution is taken from the 15^{th} step of the simulation described in Section 4.2. We can see that the proportion of exogenously activated users is independent of their endogenous activation probability $p_i(t)$. On the other hand, the higher the $p_i(t)$ the higher the proportion of endogenously activated users among all inactive ones. Vertical dashed line shows the average endogenous activation probability ($\mu(t)$ from Equation 4.2) for all inactive users. The rule for distinguishing users that activated due to endogenous or exogenous influence (Equation 4.2) classifies all newly activated users as endogenous if their $p_i(t)$ is *lower* than the $\mu(t)$, and exogenous otherwise. We see that this rule is sensible because majority of users with low $p_i(t)$ is indeed endogenously activated. Note that the baseline measure classifies users as exogenous activated if and only if $p_i(t) = 0$, which means that an user has no active peers which could exert influence on him.

4.3 Maximum likelihood method for joint inference of endogenous and exogenous influence

Maximum likelihood inference involves optimizing a log-likelihood function (Equation 3.7) conditioned on a particular forms of endogenous and exogenous influence models and the observed data - social networks structure and activation cascade. Figure 4.3b shows the visualization of a log-likelihood function in the case of a simulated activation cascade. Ideally, we would want to infer parameters of endogenous and exogenous influence for each user separately and for each time step separately. There is no reason to assume that all users have the same susceptibility to social or exogenous influence, or that this susceptibility does not change in time. However, in order to perform inference in case when only a single activation cascade is available we have to introduce some simplifying assumptions. First, we assume that the parameters of both endogenous and exogenous influence models are equal for all users at any given time. Second, parameters of endogenous influence model do not change in time while the parameters of exogenous influence may change in time. Formally, this means that our inference should result in a single set of endogenous influence parameters p^{peer} and a set of exogenous influence parameters

$\{p^{ext}\}_t$ for each time window $[t+\Delta t]$ which we use to define which users did or did not activate in a given time period. This makes the dimensionality of the final influence model (and with it the log-likelihood) proportional to the number of time windows we use for the inference - $t+1$-dimensional in the case of SI model, and $t+2$-dimensional for the EXP and LOG models.

Optimizing a model where number of parameters depends on the number of time windows that are considered in inference is not a desirable property, as such a high number of parameters makes a direct optimization unfeasible. Instead, we decided to use an *alternating* method [35] where we alternatively fix either p^{peer} or $\{p^{ext}\}_t$ and optimize for the other until both values converge.

Although this alternating method bears some similarity to the Expectation-Maximization method, it is not, in the strictest sense, an EM method because it lacks both an explicit expectation step and latent variables. In the traditional EM, one iterates between optimizing a small number of parameters of interest and calculating expectations (conditioned on the current optimized values) for the values of latent variables.

Algorithm 4.1 describes the exact alternating procedure for inference of p^{peer} and $\{p^{ext}\}_t$. In brief, it calculates first p^{peer} and p^{ext} for every time window, which then serve as initial values for the alternating procedure. In the first step, we optimize for a single set of endogenous parameters p^{peer}, conditioning on the exogenous parameters $\{p^{ext}\}_t$ we obtained for each time window. In the second step, we optimize exogenous parameters for each window separately $\{p^{ext}\}_t$, conditioning on a single set of endogenous parameters p^{peer} we obtained in the previous step. We then alternate between the first and the second step until values for p^{peer} and $\{p^{ext}\}_t$ converge.

Algorithm 4.1 Alternating method for joint inference of influence

procedure AlternatingInference(T,ϵ,$p_{peer}(t)$,$p_{ext}(t)$)
  for $t \in \{1,...,T\}$ **do**
    $\{p_{peer}\}_t, \{p_{ext}\}_t \leftarrow \text{MAP}(p_{peer}(t),\, p_{ext}(t))$ ▷ Optimize for every time window.
  end for
  while $\Delta_{peer}^{(i-1)} \geq \epsilon$ & $\Delta_{ext}^{(i-1)} \geq \epsilon$ **do** ▷ Until p_{peer} and $\{p_{ext}\}_t$ converge.
    $p_{peer}^{(i)} \leftarrow \text{MAP}(\{p_{ext}^{(i-1)}\}_t)$ ▷ Fix $\{p_{ext}^{(i-1)}\}_t$ and optimize for single $p_{peer}^{(i)}$.
    $\{p_{ext}\}^{(i)} \leftarrow \text{MAP}(p_{peer}^{(i)})$ ▷ Fix $p_{peer}^{(i)}$ and optimize $\{p_{ext}^{(i)}\}_t$ for every window.
    $\Delta_{peer}^{(i)} \leftarrow p_{peer}^{(i)} - p_{peer}^{(i-1)}$
    $\Delta_{ext}^{(i)} \leftarrow \sum_{t=1}^{T}(p_{ext}^{(i)}(t) - p_{ext}^{(i-1)}(t))$
    $i \leftarrow i+1$
  end while
  return $p_{peer}^{(i)}, \{p_{ext}^{(i)}\}_t$ ▷ The parameters of endogenous and exogenous influence.
end procedure

Procedure MAP (Maximum a Posteriori) in Algorithm 4.1 is an actual optimization

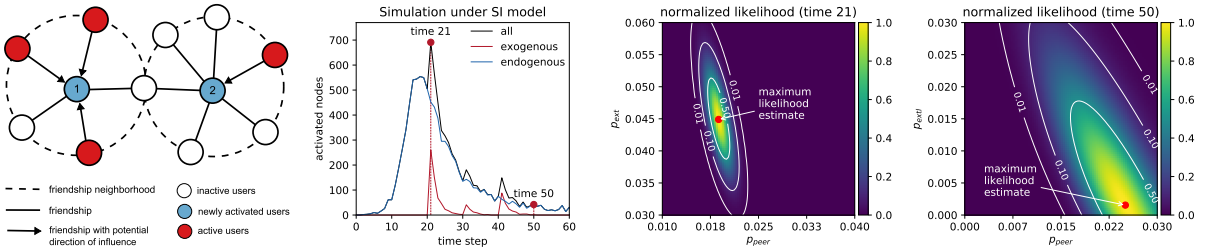
method which searches for the maximum-likelihood parameters of interest with respect to the given fixed parameters and observed data. The implementation used in these experiments uses a truncated Newton algorithm [129] which is Hessian-free - it does not require a gradient function in an explicit closed-form. It uses conjugate gradients for parameter updates in iterative fashion. The inner solver runs for only a limited number of iterations (it is *truncated*) so it is suitable for problems with large number of parameters. It also works with constrained parameters which we exploit extensively because many of our parameters are constrained in the interval $[0, 1]$ - all those that represent probabilities such as p_0 in the endogenous influence models (Equations 3.2 and 3.3), as well as all t exogenous influence parameters $p_{ext}(t)$ (Equations 3.6 and 3.7). The specific implementation of the truncated Newton algorithm that we use is from the scikit-learn Python package [†].

Convergence of this alternating method is not guarantied. The method sometimes fails to converge, especially in cases when we use the two-parameter EXP model (Equation 3.3, where parameters are p_0 and λ), or LOG model (Equation 3.4, where parameters are k and a_0). With the one-parameter SI model (Equation 3.2 that has just a single parameter p_0) inference converges without a problem, so a common trick [35] is used to reduce the two-parameter endogenous models into a single parameter model. One way to do this is by choosing several reasonable values for the parameter we want to remove, and optimizing log-likelihood multiple times by conditioning on each of these values separately. We can then choose among these the parameter value which yielded the best log-likelihood.

The output of the alternating method in Algorithm 4.1 are parameters of endogenous influence p_{peer} and exogenous influence $\{p_{ext}\}_t$. We use these inferred parameters to calculate two additional measures: 1) absolute number of activated users at each time period, and 2) type of user activation. In order to do this we first have to calculate probabilities of endogenous $p^{(i)}_{peer}(t)$ and exogenous $p^{(i)}_{ext}(t)$ activation for each individual user i at the time of their activation t. Because exogenous influence acts on all users equally regardless of the current state of their peers we can interpret exogenous influence $\{p^{ext}\}_t$ directly as the probability of exogenous activation for each user that activated at some time t. As for the probability of endogenous influence $p^{(i)}_{peer}(t)$ for a specific user i at time t we can use one of the appropriate Equations 3.2-3.4 from Section 3.3, depending on which endogenous model of influence is used in the inference. These equations depend on the activation states of all of the peers in the user's immediate social network neighborhood, at the time of his own activation.

Absolute number of activated users due to endogenous $A_{peer}(t)$ and exogenous $A_{ext}(t)$ influence are calculated by summing respective activation probabilities and normalizing with the number of users activated in a given time period, which is an observed quantity:

[†] `https://scikit-learn.org/stable/`



(a) **(b)**

Figure 4.3: Maximum likelihood inference of endogenous and exogenous influence in simulated activation cascade. Panel 4.3a shows a simple example of two newly activated users u_1 and u_2 which, at the time of activation, had different number of already active peers - user u_1 has three while user u_2 had only one. Intuitively, we would expect that user's u_1 activation is easier explained by the endogenous influence while user's u_2 is easier explained with exogenous influence. This illustrates different assumptions on the endogenous and exogenous influence used in the inference - endogenous (peer) influence acts between the users of the social network while exogenous influence is external to it. Panel 4.3b shows the simulated activation cascade using SI endogenous influence model and a designed exogenous influence with several exponentially-decaying spikes of activity (Panel 4.3b, left). The likelihood function in this case consists of only two parameters at each time step - a parameter for endogenous influence p_{SI} and a parameter of exogenous influence p_{EXT}. Estimation of these parameters separately for each time step corresponds to the initialization step in the alternating inference method (Algorithm 4.1). Visualization of the normalized likelihood function at two distinct time steps (Panel 4.3b, middle and right) shows that the two parameters are correlated - each provides a partial explanation for the observed data and if one is weaker the other must compensate. Also, when there is more data available for inference (time step 21, Panel 4.3b, middle) the shape of the log-likelihood is more concentrated around the maximum likelihood value than when there is less data (time step 50), resulting in more confident estimates. Maximum likelihood solution is obtained by optimizing a log-likelihood function with a truncated Newton algorithm [129].

$$P = \sum_{i \in \text{inactive at } t} p_{peer}^{(i)} \quad , \quad E = \sum_{i \in \text{inactive at } t} p_{ext}^{(i)} \tag{4.3}$$

$$A_{peer}(t) = |A(t)| \frac{P}{P+E} \quad , \quad A_{ext}(t) = |A(t)| \frac{E}{P+E} \tag{4.4}$$

Where $\mathsf{A(t)}$ is the number of users that activated during a particular time period $[t - \Delta t, t]$, that is $|A(t)| = |\{i \in \text{activated at } [t-\Delta t, t]\}|$ in the continuous time case (Chapter 5) and $|A(t)| = |\{i \in \text{activated at } t\}|$ in the discrete case (Section 4.5).

Type of user activation is expressed through a single measure of *exogenous responsibility* $R^{(i)}$ which quantifies to what degree is an activation of user i due to the exogenous influence:

$$R^{(i)}(t) = \frac{p_{\text{ext}}(t)}{p_{\text{ext}}(t) + p_{\text{peer}}^{(i)}(t)} \tag{4.5}$$

Here, t is the activation time of user i. Values of $R^{(i)}(t)$ are in range from 0 to 1, with close to zero indicating dominating endogenous influence, and values close to one indicating dominating exogenous influence. Users who activated during time when there was no exogenous influence acting in the social network will have $R^{(i)}(t) = 0$. On the other hand, users who, at the time of their activation, did not have any already activate peers will have $R^{(i)}(t) = 1$. It is not possible for both $p_{\text{ext}}(t)$ and $p_{\text{peer}}^{(i)}(t)$ to be 0, and consequently that the value of responsibility is undefined, because that would mean the activation of this user is evaluated as *impossible* by the model in Equation 3.7. Later in the chapter we show several alternative definitions of exogenous responsibility.

4.4 Correction for the observer bias in joint inference of influence

The Facebook friendship networks which were collected through the survey application contain only friendships between users that eventually registered on one of the survey applications. Due to the Facebook's privacy policy it is only possible to retrieve friendship relation between users that *both* registered on the application eventually [‡]. This causes the overestimation of exogenous influence, especially near the end of the data collection period where majority of users that will eventually register already did so. We call this effect an *observer bias* and it arises because the number of inactive users in the observed

[‡]As we explained in Chapter 2, Facebook's privacy policy and API for data collection changed several times over the years, usually in the direction which restricted type and amount of data which could be collected by third party application developers. This particular change was introduced in early 2014.

social network is getting smaller and smaller, while in actual network (being much larger than what we observe) their number does not change much. To correct for this bias we artificially extended the social network with inactive users by a certain fraction α of the total number of registered users. The α is introduced through a correction factor $c(t)$:

$$c(t) = 1 + \alpha \frac{N_{\text{all}}}{N_{\text{inactive}}(t)} \tag{4.6}$$

This correction factor can be included in the log-likelihood function to modify the part responsible for the inactive users:

$$\begin{aligned} \log \mathcal{L}(D;\, p_{peer}, p_{ext}, t) = & \sum_{i \in \text{activated at } [t-\Delta t, t]} \log(1 - (1 - p_{peer}^{(i)}(t))(1 - p_{ext}(t))) + \\ & c(t) \sum_{i \in \text{inactive at } t} \log((1 - p_{peer}^{(i)}(t))(1 - p_{ext}(t))) \end{aligned} \tag{4.7}$$

In the case of $\alpha = 0$ we are not making any correction for the observer bias at all, and we can expect to overestimate exogenous influence near the end of the observation period.

4.5 Joint inference of endogenous and exogenous influence on simulated data

We test the method for joint inference on simulated activation cascades which are designed to be as similar as possible to condition present in empirical data. As social network we use a configuration model of a Facebook friendship network collected through referendum2013 survey application. Configuration model preserves the degree of each node - number of Facebook friends each user has, but permutes their mutual connections. The degree sequence, number of users and total number of connections are preserved, but mesoscale network structures such as communities are not. However, this is still more preferable than completely permuting the connections without preserving the degrees of nodes, because this effectively changes the degree distribution to binomial [130]

The simulated activation cascade is initialized with a small number of activated user and progresses in discrete time steps. Both endogenous and exogenous influence are active at the same time. Endogenous influence is modeled with one of the endogenous influence models which define influence that active users exert on inactive users (Section 3.3, Equations 3.2-3.4). In the simulation in Figure 4.4 we use EXP model, but the results for other endogenous influence models are similar (Figure 4.5). Exogenous influence is modeled with a designed non-parametric influence which acts equally on all inactive users. For this simulation we had chosen a shape that features three exponentially-decaying peaks of

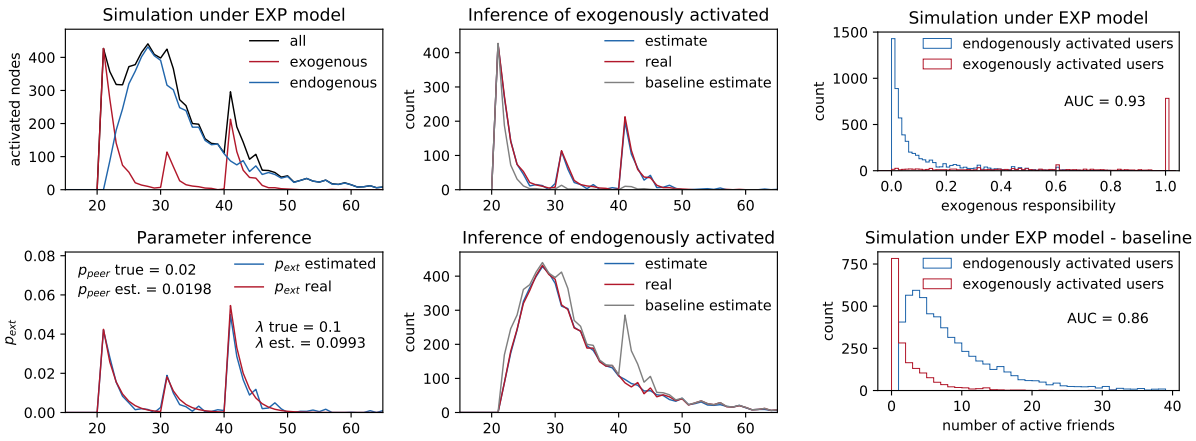
exogenous influence, which resembles a typical situation when a distinct exogenous information source activates some of the users [131], which is a pattern that we also observe in the collected dataset (Figure 2.5). However, the proposed inference method works equally well for various other shapes of exogenous influence (Figure 4.6).

The input to the inference method are the activation times of all users - a single activation cascade, and a Facebook friendship network between users - an online social network. Using just these two information and by employing alternating inference method in Algorithm 4.1 we can estimate the parameters of the assumed endogenous and exogenous influence models as well as the absolute number of users activated due to one of these influences.

A measure of *external responsibility* $R^{(i)}(t)$ (Equation 4.5) gives us an estimate to what extent was user's activation due to endogenous or exogenous influence. If we wanted to classify users as being endogenously or exogenously activated we could use a specific threshold for $R^{(i)}(t)$, and compare this with the ground truth data which is available in the simulation. However, we decided to instead evaluate the estimate across the whole range of possible thresholds, and so we calculated the whole receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) measure to evaluate the performance (Figure 4.4b, Panel 4.4b). The achieved AUC is 0.93 which indicates significant predictive performance. Similar as in simulated experiments in Section 4.2, we again compare the proposed method to a simple baseline measure commonly used in previous work [35, 120] where an activation is considered exogenous if activated user had no other active peers at the time of activation. This baseline underestimates the number of exogenously activated users, especially near the end of the data collection period when it becomes increasingly likely that an user will be connected to at least one active peer due to chance rather than to some kind of social influence between them.

The inference is fast and scales well to social networks of over ten thousand users (Section 4.6).

Similar results are obtained for other endogenous influence models, namely the SI (Figure 4.5a) and the LOG model (Figure 4.5b). Similarly as when using EXP as the endogenous influence model, the alternating inference method correctly infers the parameters of both endogenous and exogenous influence models, absolute number of users activated due to endogenous or exogenous influence, as well as characterize the activation of each user as being driven dominantly by one or the other influence. The achieved AUC is 0.92 for SI model 0.94 for LOG model, which is similar to the results obtained with the EXP model. This shows that the alternating inference method in Algorithm 4.1 can be used both for simple influence models such as SI and EXP where each user has an independent probability of activating any of his peers, as well as for complex influence models [4]



(a) **(b)**

Figure 4.4: Joint inference of endogenous and exogenous influence on a simulated activation cascade using EXP endogenous influence model. Social network is a configuration model of a Facebook friendship network collected through referendum2013 survey application. Assumed endogenous influence model is EXP, and exogenous influence is designed manually to feature three distinct peaks of exogenous activity. Input to the inference procedure are user activation times (black line in Panel 4.4a) and a friendship network between users. The alternating inference method (Algorithm 4.1) is able to estimate the absolute number of endogenously and exogenously activated users throughout the whole simulation period and to correctly infer the parameters of endogenous influence - p_{peer} and λ, and exogenous influence $\{p_{external}\}_t$ which is defined for every time period $[t + \Delta t]$. We also infer activation type for each user individually by using the exogenous responsibility measure $R^{(i)}(t)$ (Equation 4.5) as shown on Panel 4.4b and achieve AUC of 0.93. In comparison, the baseline method (gray line, Panel 4.4a), underestimates the number of exogenously activated users, especially near the end of the simulation. Baseline method on Panel 4.4a is a special case where users are defined as exogenously activated if they did not have any active peers at the time of activation. We can loosen this criteria (similar to what we did in Equations 4.1 and 4.2) and calculate AUC across all possible values of the number of active peers to asses this measure's utility for classification of users into endogenously and exogenously activated. In this case we achieve AUC of 0.86, which is lower than what is achieved with the alternating inference method. Even by just observing the histograms of the two measures on Figure 4.4b we see that the $R^{(i)}(t)$ is much better in differentiating the two types of users.

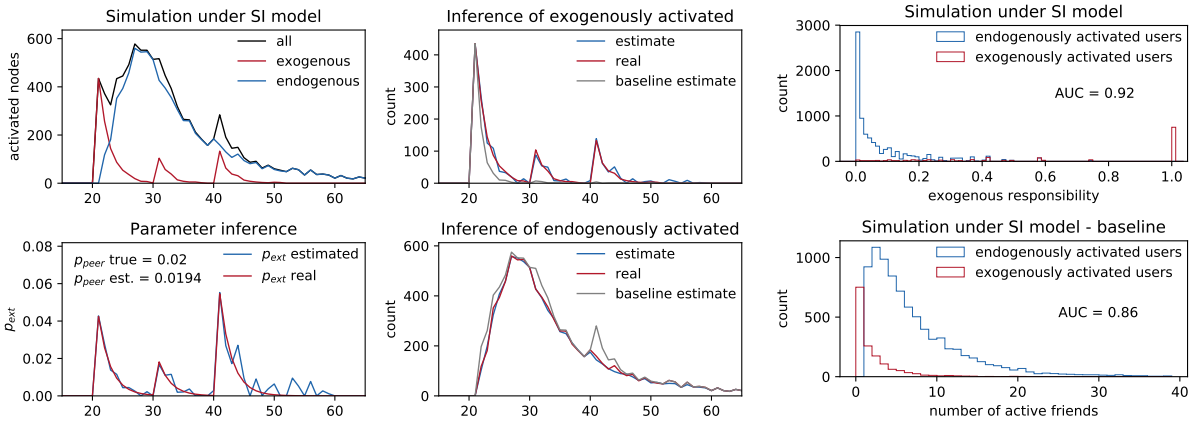
such as LOG the probability of activation depends on the aggregate property of the user's neighborhood in the social network. Instead of the before mentioned endogenous influence models we could, in theory, use any other microscopic model of endogenous influence which can be efficiently computed given information on user's friendship connections and the activation state of his peers.

The previous experiments on simulated activation cascades (Figures 4.4 and 4.5) used a spike-shaped exogenous influence. This type of shape mimics a common situation in empirical data where we have a surge in user activity following an external event which then decays gradually in time. In our case these external events are often related to the publication of the online news media articles which provided a link to the survey application (Figure 2.5). In general case they may correspond to any external real-life event. For example, similar spikes of user activity are also observed in Google search queries related to sudden catastrophic events [131]. The inference methodology handles exogenous influence *non-parametrically* - at each time step individually, and so does not impose any restrictions on its functional form. It can easily handle exogenous influence with arbitrary time-varying shape, including constant, exponentially decaying, sinusoidal, rectified or any combination of these (Figure 4.6). This is true even for exogenous influence which has a functional dependency on some dynamical property of the social network or the activation cascade, for example the number of currently active users. In theory, any function could be used for modeling exogenous influence, given that it is possible to calculate its value at each time step.

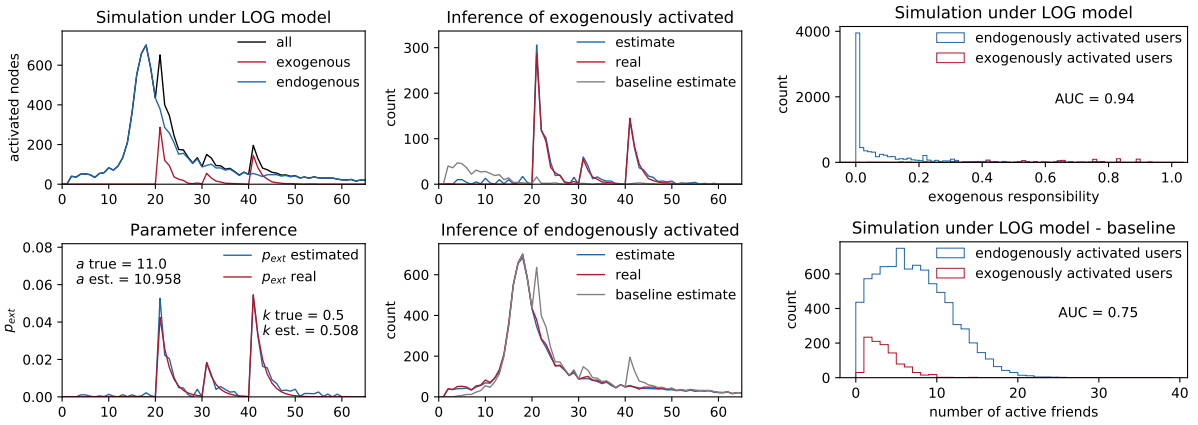
Along with the original formulation of exogenous responsibility in Equation 4.5 we can also use several other formulations. These differ in the way that exogenous activation probability $p_{\text{ext}}(t)$ and endogenous activation probability $p^{(i)}_{\text{peer}}(t))$ are aggregated to achieve a single measure of exogenous influence $R^{(i)}(t)$ for each user i at the time of its activation t. Figure 4.8a shows the original formulation as a function of endogenous and exogenous activation probabilities across the whole range $[0, 1]$. The second definition is a softmax version of the original formulation (Figure 4.8b):

$$R^{(i)}_{\text{softmax}}(t) = \frac{\exp(p_{\text{ext}}(t))}{\exp(p_{\text{ext}}(t)) + \exp(p^{(i)}_{\text{peer}}(t))} \tag{4.8}$$

The softmax formulation has the property that high values of exogenous responsibility $R^{(i)}(t)$ *cannot* be achieved for low values of endogenous activation probability, which is not true for the original formulation (Figure 4.8a). The third formulation of $R^{(i)}(t)$ (Figure 4.8c) defines it as a probability that user i activated due to exogenous influence but *not* due to endogenous influence at time t:



(a)



(b)

Figure 4.5: Joint inference on simulated activation cascades using SI (Panel 4.5a) and LOG (Panel 4.5b) models for endogenous influence, which are representative examples of *simple* (SI) and *complex* (LOG) social influence models. For both simulations we used the same shape of exogenous influence as the one in Figure 4.4, with three distinct exponentially decaying spikes of influence which resemble typical situation encountered in empirical activation cascades (Figure 2.5). The results demonstrate that the alternating inference method in Algorithm 4.1 is able to infer parameters for both endogenous and exogenous influence models, as well as estimate to what extent is the activation of each individual user due to the one or other influence. The achieved AUC scores using the measure of exogenous responsibility $R^{(i)}(t)$ are 0.92 and 0.94 in simulations using SI and LOG endogenous influence models respectively. In comparison, baseline method, which uses the number of active peers as the classification criteria, achieves AUC score of 0.86 and 0.75 for the SI and the LOG models respectively.

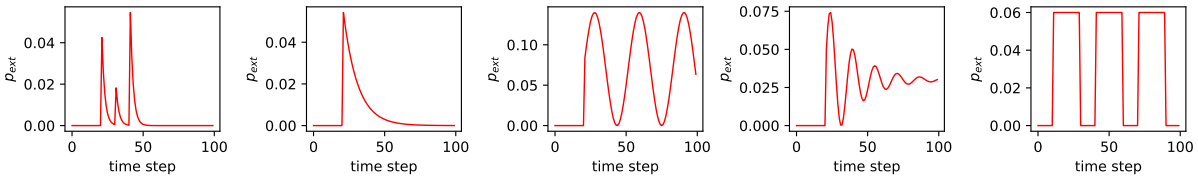


Figure 4.6: Different time-varying shapes of exogenous influence that are used in simulations (Figure 4.7). We used (from left to right) a typical exponentially-decaying shape with multiple peaks and with a single peak, sinusoidal, decaying sinusoidal and a rectified shape. As the exogenous influence is evaluated non-parametrically - at each time step, the only requirement for its use in the alternating inference method (Algorithm 4.1) is that its value can be calculated for each time step. The following shapes are independent of the current state of social network or the activation cascade, although such functional dependency could easily be included.

$$R^{(i)}_{\text{multiply}}(t) = p_{\text{ext}}(t)(1 - p^{(i)}_{\text{peer}}(t)) \tag{4.9}$$

Similar as in the softmax formulation (Equation 4.8), high values of exogenous responsibility $R^{(i)}(t)$ *cannot* be achieved for low values of endogenous activation probability. Qualitatively, all three formulations in Equations 4.5-4.9 satisfy the requirement that the larger the p_{ext} is, the larger the exogenous responsibility. The difference between the original formulation in Equation 4.5 and formulations in Equations 4.8 and 4.9 is that the former calculates $R^{(i)}(t)$ in *relative* terms - even small values of exogenous activation probability p_{ext} can achieve high exogenous responsibility if endogenous activation probability is accordingly small. This is a desirable property because it captures a *relative* rather than *absolute* difference between endogenous and exogenous activation probability. Figures 4.8a-4.8c show this property visually. To confirm that the original formulation in Equation4.5 is really the most sensible one we repeated the experiment with simulated activation cascade using SI model of endogenous influence on Figure 4.5a. We then compared AUC scores obtained by using each of the three exogenous responsibility formulations. The original one, in Equation 4.5 and Figure 4.8a, achieved the best AUC score of 0.92. The other two formulations achieved lower AUC scores - softmax formulation in Equation 4.8 and Figure 4.8b achieved AUC of 0.88 and the formulation in Equation 4.9 and Figure 4.8c achieved AUC of 0.89.

As an alternative to measures of exogenous responsibility in Equations 4.5-4.9 we can use normalized exogenous activation probability $p^{(i)}_{\text{ext}}$ directly. It may seem that we are disregarding the value of endogenous influence in this way completely - however, it is already implicitly accounted for because the values of endogenous and exogenous influences are jointly estimated from the alternating inference procedure in Algorithm 4.1. This is visible form the shape of the likelihood function (Equation 3.6) in Figure 4.3 which shows that the parameters for endogenous and exogenous influence are coupled - if one is

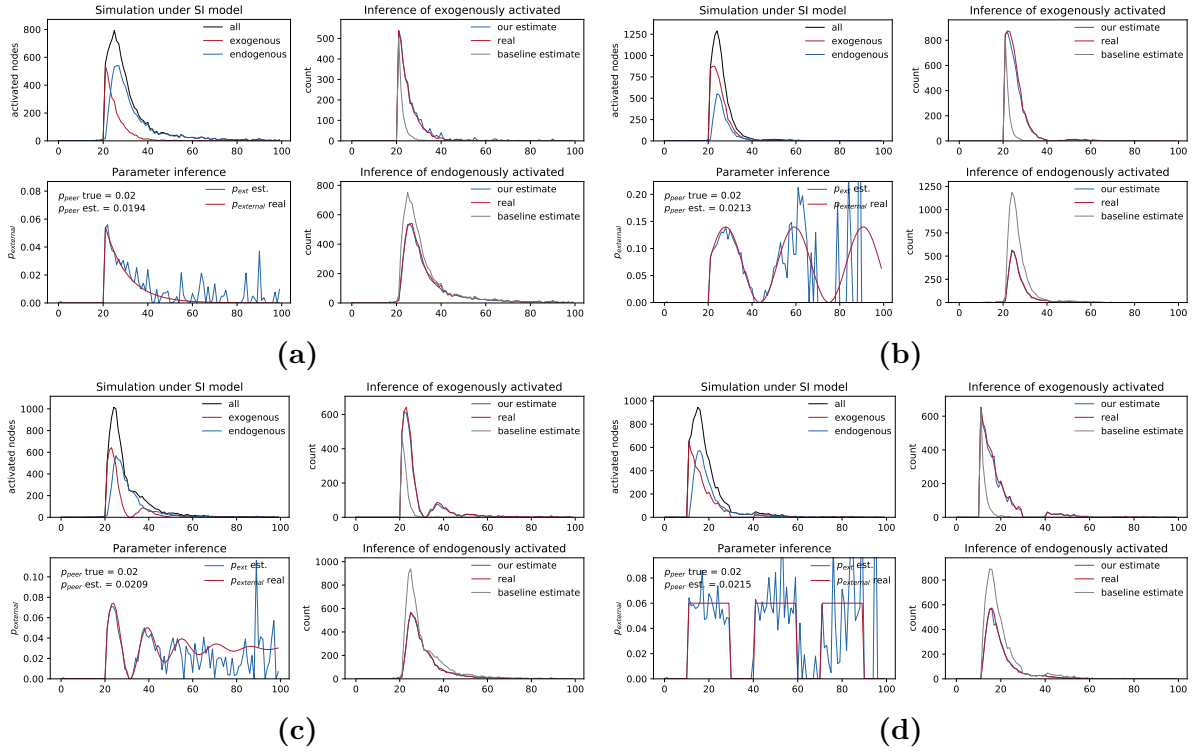
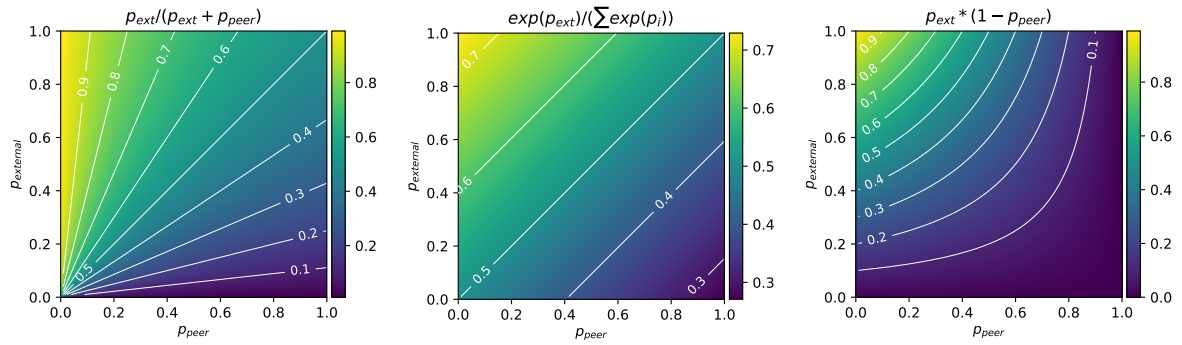


Figure 4.7: Simulated activation cascades which use different shapes of exogenous influence in Figure 4.6. Regardless of the specific shape of exogenous influence the alternating inference method (Algorithm 4.1) is able to infer parameter values of both endogenous and exogenous influence and absolute number of users activated due to the one or the other influence in all cases. We again show a comparison with the baseline method which classifies users as exogenously-activated only if they had no active peers at the time of their own activation. Similar as in experiments in Figures 4.5 and 4.5, this baseline is too conservative and underestimates the number of exogenously activated users, especially near the end of the observation period.

weaker the other one has to be stronger in order to compensate. We repeat the experiments in Figures 4.4 and 4.5, but this time using exogenous activation probability $p_{\text{ext}}^{(i)}$ instead of exogenous responsibility $R^{(i)}(t)$ as a measure of exogenous influence. Figure 4.9 compares the histograms and ROC curves obtained shows the comparison when using exogenous activation probabilities as opposed to exogenous responsibility with each of the three endogenous influence models in simulated activation cascades. First observation is that using exogenous activation probability directly results in coarser estimates of the influence, as the underlying exogenous influence curve we use assumes just a few distinct values during short time before it dissipates due to exponential decay. Also, achieved AUC scores are typically $2-8\%$ higher when using exogenous responsibility than when using endogenous activation probability directly.



(a) **(b)** **(c)**

Figure 4.8: Different formulations of exogenous responsibility $R^{(i)}(t)$. Qualitatively they all satisfy the same requirement - the larger the exogenous activation probability p_{ext} the larger the exogenous responsibility. But only the formulation in Panel 4.8a (Equation 4.5) satisfies a property that the high values of exogenous responsibility can be achieved with a relatively low values of p_{ext}. This makes sense in our case because we want the exogenous responsibility to reflect a *relative* difference between p_{ext} and p_{peer} rather than an absolute difference. To confirm this we repeat the experiment on Figure 4.5a where we simulate activation cascades under SI endogenous influence model, and compare three different formulations of exogenous responsibility. The best AUC of 0.92 is achieved using original formulation from Equation 4.5 (Panel 4.8a). The other two formulations achieve 0.88 (Equation 4.8, Panel 4.8b), and 0.89 (Equation 4.9, Panel 4.8c). This indicates that the original formulation is indeed the most appropriate.

4.6 Scalability of inference

In order to test the scalability of the alternating method in Algorithm 4.1 we repeat the inference experiments using SI endogenous influence model (Figure 4.5a) and EXP endogenous influence model (Figure 4.4a), but using increasingly larger social networks. We construct the artificial social networks by using Holme and Kim algorithm for generating networks with powerlaw degree distribution and desired average clustering, implemented in `powerlaw_cluster_graph` [§] function in NetworkX Python library. Algorithm iteratively adds nodes to network, one by one, each with three new edges (parameter `m=3`) which are then preferentially attached to the already present nodes depending on their degree. Preferential attachment produces social networks with power-law degree distribution, a property which is also observed in the collected Facebook social network dataset (Figure 2.3). Algorithm also adjusts the clustering coefficient of the generated network, which is also a property commonly found in empirical social networks. In these experiments the clustering probability is set to 0.1 (parameter `p=0.1`). We explore network sizes ranging from 10 to 1000 (Figure 4.10). Execution times are almost linear with respect to the size of the networks on which inference is being done. The inference was run on a

[§] `https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.generators.random_graphs.powerlaw_cluster_graph.html`

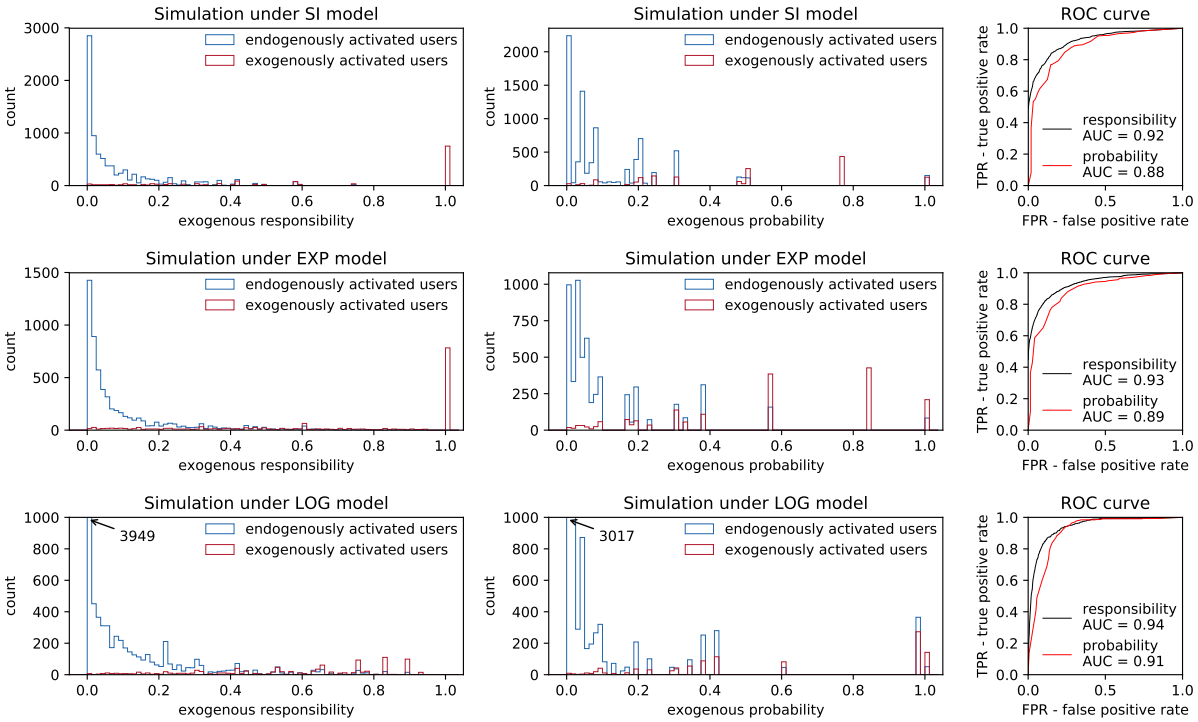


Figure 4.9: Inference on simulated activation cascades using exogenous activation probability $p_{\text{ext}}^{(i)}$ instead of exogenous responsibility $R^{(i)}(t)$ (Equations 4.5-4.9) for quantifying the magnitude of exogenous influence for each user i at the time of his activation t. Experiments are equivalent to those in Figures 4.4 and 4.5 and are using three different endogenous influence models - SI (top panel), EXP (middle panel) and LOG (lower panel). Estimates obtained with exogenous activation probability $p_{\text{ext}}^{(i)}$ alone are coarser - they exhibit less diverse values than when using exogenous responsibility $R^{(i)}(t)$ due to the fact that underlying exogenous influence curve weakens sharply in time after each of the exponentially decaying peaks. This is more pronounced for high values of exogenous activation probability. The AUC scores are typically 2% to 8% worse than when using the exogenous responsibility measure from Equation 4.5.

64-bit Intel i5-2500 CPU 3.3 GHz and 8 GB of RAM, Python 3.6.1. as a part of Anaconda distribution.

4.7 Individual and collective influence of users

Estimates of endogenous $p_{peer}^{(i)}$ and exogenous $\{p_{external}\}_t$ influence which we obtain by alternating method in Algorithm 4.1 can be used to infer to what extent is each user responsible for activation of all of his peers which activated after him. That is - $t_i < t_j$, where t_i and t_j are activation times of users i and j respectively. This is what we call an *individual influence*. By aggregating this individual influences across a set of users we can estimate their *collective influence*. An underlying assumption is that users can claim only the endogenous part of their peer's activation, as exogenous activation is beyond each user's control. However, our estimates of these two influences are not a substitute for a deterministic activation path - we do not know exactly who shared information with

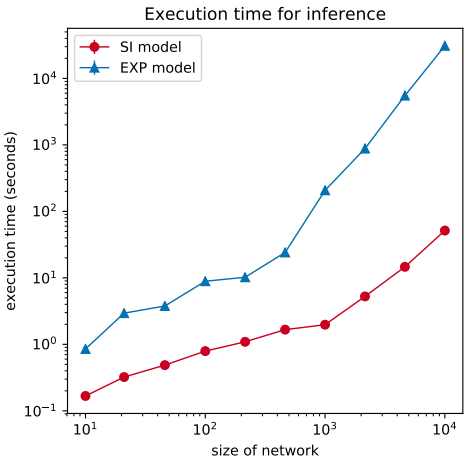


Figure 4.10: Scalability analysis for the inference methodology is performed by running alternating inference method in Algorithm 4.1 on simulated activation cascades using SI and EXP endogenous influence models and the designed exogenous influence curves from experiment in Figures 4.5a and 4.4a. We explore generated social network ranging from 10 to 1000 to users. Execution times are almost linear with respect to the size of the networks.

whom. This is why we cannot compute individual influences in a transitive fashion by counting all peers along the activation path [132]. What we do instead is to express the influence $I^{(i)}$ of user i (Equation 4.10) as the extent to which user i *might* be responsible for activation of each of his peers $N^{(i)}$:

$$I^{(i)} = \sum_{j \in N^{(i)}} \frac{I^{(i \to j)}}{\sum_{m \in N^{(j)}} I^{(m \to j)}} p^{(j)}_{peer} \tag{4.10}$$

Here, the quantity $I^{(i \to j)}$ is the fraction of endogenous influence that user i can claim for the activation of his peer j. In our case this is simply $I^{(i \to j)} = 1$, but in general this expression could be arbitrarily complex. For example, we could make it dependent on time t, where the rationale is that users can claim higher fraction of their peer's endogenous activation if they activated *closer* in time. If we make this time dependence exponentially-decaying, that is $I^{(i \to j)} = e^{-\lambda(t_j - t_i)}$, the expression for individual influence becomes:

$$I^{(i)}_{\text{EXP}} = \sum_{j \in N^{(i)}} \frac{e^{-\lambda(t_j - t_i)}}{\sum_{m \in N^{(j)}} e^{-\lambda(t_j - t_m)}} p^{(j)}_{peer} \tag{4.11}$$

Another option is to make $I^{(i \to j)}$ dependent on some intrinsic characteristic of user itself. In this way we could encode a requirement that similar users (for example, those that belong to the same community) are more probable of influencing each other. The choice of which expression to use, whether that from Equation 4.10 or from Equation 4.11, is independent on the particular type of endogenous influence model used. In experiments

described in this thesis only the simplest definition in Equation 4.10 is used for all endogenous influence models defined in Equations 3.2-3.4. As it is already mentioned, because we do not have a deterministic activation path the $I^{(i \to j)}$ really encodes just a potential for endogenous influence. In practice any of j's peers could equally be responsible for his activation, not just the user i. This is why the denominator of Equation 4.10 has a normalizing term for all other peers k of user j. In the case of SI model this assigns to each user $1/k$ of the peer activation probability $p_{peer}^{(j)}$ for each of his peers j that activated after him ($t_i < t_j$), where k is the number of user's j peers that activated before him.

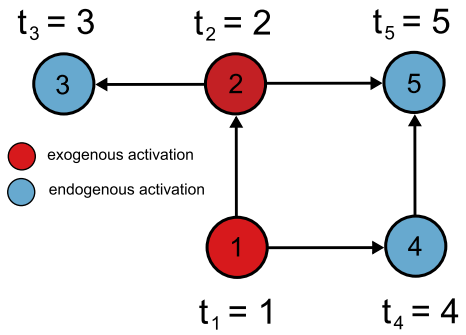
The collective influence for a group of users G is just an average influence of all users in the group $1/\mathsf{G} \sum_{i \in G} I_i$.

Illustrative example of the calculation of individual influence is shown bellow. We start with a social network consisting of five users $\{u_1, u_2, u_3, u_4, u_5\}$. Their friendship connections are encoded with an adjacency matrix A. We also have activation time $\{t_i\}$ of each user and the type of activation - whether it was endogenous or exogenous. The arrows on the schematic illustration bellow indicate a potential direction of endogenous influence between users that have a friendship connection. If there are two users u_i and u_j of which user u_i activated before user u_j, that is $t_i < t_j$, the arrow will point from user u_j towards user u_i.

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$\text{activation times} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

$$\text{endogenous activation} = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$



The type of activation is usually not directly observable in data, so we either have to estimate it with an inference method like ours or use a proxy, for example, referral links which encode user's digital traces. Before we continue we need to calculate several quantities for each user individually - i) number of peers active *at* each user's activation, and ii) number of peers which activated *after* each user's activation:

$$\text{number of active peers } \textit{at} \text{ activation} = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

$$\text{number of active peers } \textit{after} \text{ activation} = \begin{bmatrix} 2 & 2 & 0 & 1 & 0 \end{bmatrix}$$

Let us explain step-by-step how to calculate individual influences users u_1 and u_2. For

user u_1, two of his peers (u_2 and u_4) activated after him, but only u_4 due to endogenous influence. User u_4 has no other peers that activated before him, so user u_1 gets the full credit for his endogenous activation, making his individual influence 1.0. User u_2 has two peers that activated after him (u_3 and u_5), and both activated due to endogenous influence. User u_3 has no other peers that activated before him, so user u_2 gets full credit for his endogenous activation, which is 1.0. User u_5 has one additional peer that activated before him (u_4) and so the credit for his endogenous activation should be equally divided by users u_2 and u_4. So user u_2 is assigned 0.5 of influence for the activation of user u_5, making his total individual influence 1.5. Note that if we used expression for individual influence from Equation 4.11, due to the term $e^{\lambda(t_j - t_i)}$ in the nominator, users u_2 and u_4 would not be assigned equal credit for endogenous activation of user u_5. In that case, because user u_4 activated closer in time ($t_4 = 4$) to user u_5 ($t_5 = 5$) than user u_2 ($t_2 = 2$), user u_4 would be assigned larger credit for the endogenous activation of user u_5. The final influence I for all users is:

$$\mathrm{I} = \begin{bmatrix} 1.0 & 1.5 & 0.0 & 0.5 & 0.0 \end{bmatrix}$$

Intuitively, we see that user u_2 is estimated to be the most influential. This makes sense because it has two peers that activated after him and both due to endogenous influence. The least influential users are u_3 and u_5 that have no peers at all.

4.8 Comparison of influence with the structural measures on simulated data

In Figure 4.11 individual influences of users calculated with Equation 4.10 are compared with several structural measures on simulated activation cascades, under similar simulation conditions as experiments in Figure 4.5a. The structural measures are defined on a user level and are: (i) number of peers which *activated* after the activation of a particular user, (ii) number of peers, (iii) activation time, and (iv) number of peers active *before* the activation of a particular user. The Spearman correlation coefficient is displayed for each of the scatter plots on Figure 4.11. We can see that the measure of influence is more correlated with the number of peers that activated after each particular user ($R = 0.93$) than with the number of peers that activated before ($R = 0.33$). This is intuitively clear as we expect that influence acts *forwards*, rather than *backwards* in time. Also, there is a positive correlation with the number of peers ($R = 0.74$) and negative correlation with the activation time ($R = -0.72$). This is also unsurprising as we expect that users with more peers have more opportunity to exert their influence, while on the other hand late

activation times means these opportunities are reduced.

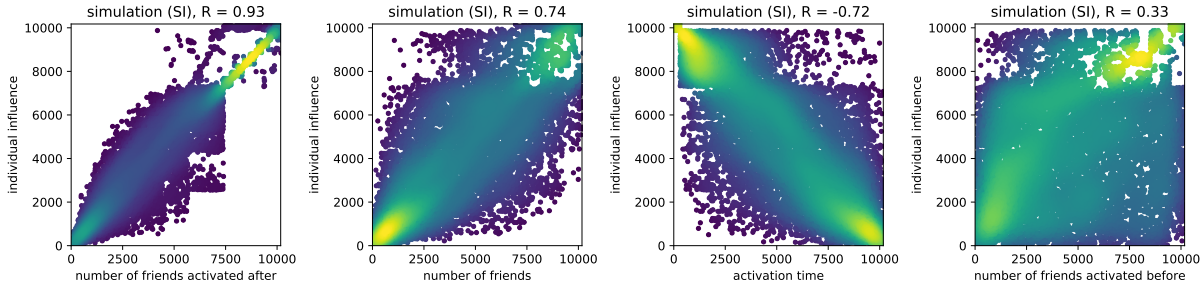


Figure 4.11: Comparison of user influence with structural measures on simulated activation cascades. Conditions of the experiment are similar to those in Figure 4.5a, where user friendship network is generated from a configuration model of the referendum2013 friendship network, and SI model is used for endogenous influence with a spike-shaped exogenous influence. Scatter plots show comparison of user's ranks of their individual influence, calculated with Equation 4.10, as compared to one of the four measures of structural influence: (i) number of peers which activated *after* the activation of a particular user, (ii) number of peers, (iii) activation time, and (iv) number of peers active *before* the activation of a particular user. Points are colored based on the density in the surrounding plot region - points in denser regions are colored yellow instead of blue. Spearman correlation coefficient R is calculated for each of the plots.

Chapter 5

Evaluation

In Chapter 4 results on simulated activation cascades where presented. These simulations tried to emulate, as close as possible, conditions which are expected to appear in empirical datasets such as those described in Chapter 2. This includes the degree distribution of simulated social networks and hypothesized shape of exogenous influence which was based on those observed in empirical data. The results on simulated data were promising, but in order to properly evaluate the inference methodology proposed in Chapter 4 we have to apply it to the actual empirical data. Evaluation on simulated data is easier because it is possible to compare the inferred parameters with the underlying influence models directly. Also, the chosen evaluation measure AUC only makes sense if the underlying types of activation - whether user was activated due to endogenous or exogenous influence, are known. These are actually not known in empirical data so a proxy measure needs to be used for evaluation of the proposed inference method.

In the remainder of this chapter two approaches to evaluation are described. A first attempt for modeling exogenous and endogenous influence [101] (Section 5.1), where a direct method for inference of endogenously and exogenously activated users is used (Section 4.2), estimates of endogenously activated users are evaluated by comparing it with the known number of users visiting the online survey application from a referral link originating in Facebook. This first empirical evaluation is performed on referendum2013 dataset (10175 users) where no referral links were collected from users. Instead, aggregated estimate from Google Analytics which tracked users visiting the survey application were used.

Second attempt [66] (Section 5.2) uses a maximum likelihood method for joint inference of endogenous and exogenous influence (Section 4.3) and estimates are evaluated on all three empirical datasets - referendum2013 (10175 users), sabor2015 (6909 users) and sabor2016 (3818 users). Instead of choosing a single threshold for estimation like in Section 5.1, the whole Receiver Operating Characteristic (ROC) curve [133] and the asso-

ciated area under the curve (AUC) score are calculated. AUC score gives the probability that the inference method gives a higher *exogenous responsibility* score to a randomly chosen exogenously activated user as compared to a randomly chosen endogenously activated user. The purpose of these evaluation measures is to provide an estimate of how well does the proposed influence model fit the empirical data, provided that the underlying assumptions of the modeling procedure are satisfied.

In order to calculate AUC score we need to know actual activation type for each user, As a proxy for this an information obtained from referral links which were collected for users of sabor2015 and sabor2016 online survey applications can be used. Based on referral links users can be categorized into one of three categories:

1. *Strong endogenous influence:* Users whose referral link originates from a Facebook share.
2. *Potential endogenous influence:* Users whose referral link originates from Facebook.
3. *Strong exogenous influence:* Users whose referral link originates from an external web site.

Users that do not have a referral link are categorized as unknown and are not used in evaluation. To actually calculate ROC and AUC score a binary decision problem needs to be defined. In the experiments presented in this chapter users from first category are considered as endogenously activated and users from third category are considered as exogenously activated.

5.1 Inference exogenous influence directly on empirical datasets

In this section the method of direct inference of endogenous and exogenous influence outlined in Section 5.1 is used on referendum2013 dataset. This dataset contains Facebook friendship relations between 10175 users and their registration times. Equation 4.1 is used for calculation expected probability of endogenous activation $\mu(t)$ over all inactive users at time t. Estimate of $\mu(t)$ can then be used in Equation 4.2 to calculate the absolute number of endogenously activated users $peer(t)$. A crucial missing component needed for the calculation of $\mu(t)$ is the probability of endogenous activation $p_i(t)$ for each user i, which is defined in Equation 3.1. For this the two endogenous influence parameters are needed - λ and p_0, which define an exponentially-decaying influence between users. These parameters are inferred from data using information on the visitors to the online survey application which were collected using Google Analytics API. Unfortunately, in this way only a total number of visitors could be collected, so we do not have a finer time resolution of their visits similar to what is collected with Facebook Graph API (Figure 2.5). Also,

with information collected through Google Analytics it is only possible to tell whether visitors came from Facebook or some specific external website, not whether they followed a Facebook share from another user. This is why it is not possible to classify users into the three categories outlined in the beginning of this chapter. Concretely, we cannot distinguish between visitors with "strong" endogenous influence (that followed a Facebook share) and "potential" endogenous influence (that came from Facebook in general).

Using information from Google Analytics, total of 25154 visitors visited the website which hosted the online survey application. Out of these, 17587 came from a referral link originating in Facebook, while the rest came from an external website, usually from online news websites which reported on the survey application. This means that the approximate ratio of endogenously activated users among all users is 70%. A fine-grained grid search is performed consisting of all (λ, p_0) parameter combinations so that the percentage of endogenously activated users, as estimated by model in Equation 4.2, during the first day of user registrations period is equal to 70%. The Figure 5.1 shows the results of parameter optimization. We identify a curve in (λ, p_0) parameter space where parameters satisfy the given constraint. For illustration purposes values of $p_0 = 0.6$ and $\lambda_p = 0.001$ are chosen as parameters of endogenous influence and are used to estimate the absolute number of endogenously and exogenously activated users on Figure 5.2. However, the estimates are robust even if different (λ, p_0) values are chosen from the given curve in the parameter space (see again Figure 5.1).

For evaluation of the estimates of endogenous and exogenous influence several proxy measures are used, as for the referendum2013 dataset there is no information on the referral links from which users visited the online survey application. So we cannot characterize each user's activation *individually* but have to characterize it *collectively* instead. As a proxy for exogenous influence publication times of online news articles which reported on the survey application during the collection period are used. Google Analytics additionally gives us a total number of visitors visiting the online survey application by following a referral link from these domains (Figure 5.2). Although this is a total visitor count - there is no any time resolution of visits, it is still useful due to the fact that majority of visits happens in a short time just after the publication of the online news article. This gives us an approximate estimate for the magnitude of the exogenous influence, as more visitors from an external website indicates a stronger exogenous influence. By observing the time series of users registrations (Figure 2.5 shows this for all three datasets) we see that after each news publication there is a sudden rise in user registrations that weakens over time. This is captured with the direct method estimates on Figure 5.2, especially during the first day of user's registrations where the method estimates two peaks in exogenous influence right after publication of news articles from two major online news portals

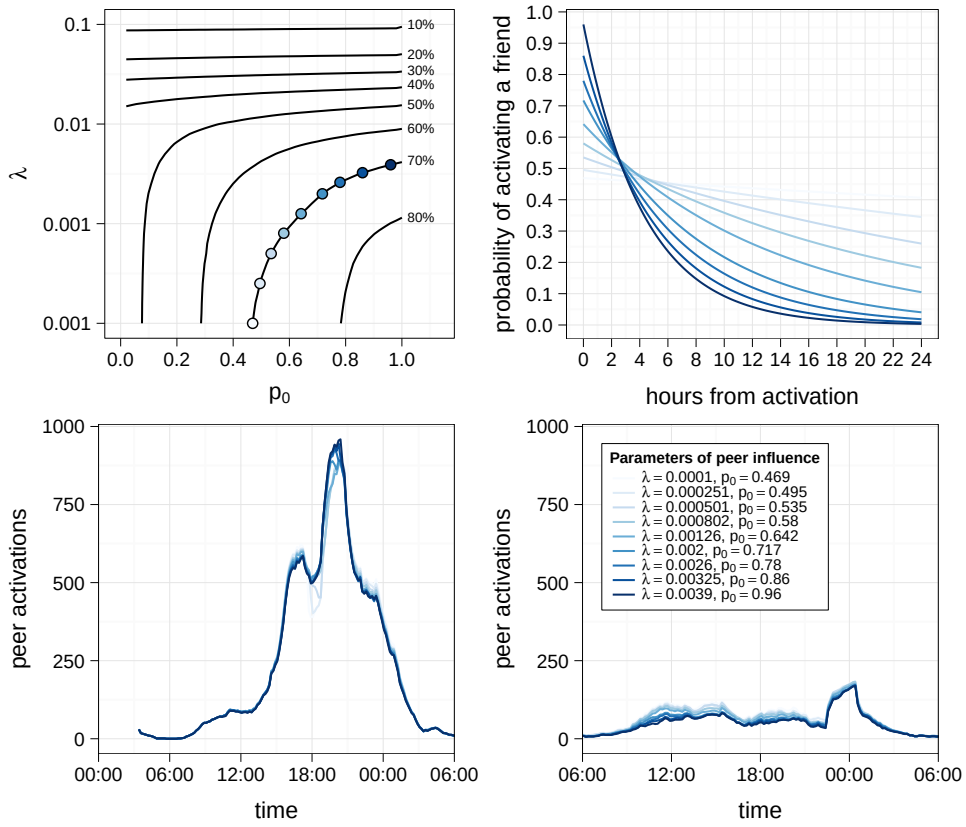


Figure 5.1: Choosing optimal endogenous influence parameters (λ,p_0) that determine exponential decay of influence (Equation 3.1). A full grid search of the parameter space shows that low values of λ and high values of p_0 correspond to the strong endogenous influence that decays slowly, resulting in a higher number of endogenously activated users (top left Figure). The values are chosen so that the fraction of endogenously activated users during the first day of data collection equals 70% - a value which corresponds to the fraction of users visiting the online survey application by following a link from Facebook as determined by Google Analytics. All parameter combinations laying on the 70% curve are optimal by this criterion. For all of these plot the influence curves (top right Figure) and the estimated number of endogenously activated user for the first and the third day of data collection (bottom Figures). Values of $p_0 = 0.6$ and $\lambda_p = 0.001$ are later used for estimating absolute number of endogenously and exogenously activated users during the whole observation period (Figure 5.2).

(vecernji.hr with 2027 total visits and jutarnji.hr with 1637 total visits).

As for the endogenous influence, again there is no information on referral links from which users visited the survey application so it is not known which particular user followed a referral link from Facebook. However, a total count of visitors is known due to the information provided by Google Analytics service. This information is already used for the choice of optimal parameters of endogenous influence (λ, p_0) (Figure 5.1). In addition, two proxy measures are used to evaluate the estimates of endogenous influence, both of which rely on certain periods of data collection where there is an additional independent information on the type of user's registrations. First, it is known that before the first online news article was published, which happened about 12 hours after referendum2013 went online, that the majority of user registrations had to be due to endogenous influence, as information on the survey spread organically between users without any major external information source. Figure 5.2 shows the estimate of exogenous influence using a direct method during the first 12 hours - it is almost negligible, and it rises just after the first news publication by vecernji.hr which brought many new users to the survey which were otherwise poorly connected to the already registered users. In comparison, the baseline method correctly identifies the first peak of exogenous registrations, but soon starts to underestimate the number of exogenously activated users by completely ignoring them after the first day, which is unrealistic. This happens because of the too conservative criterion which baseline puts on exogenously activated users - it classifies as exogenous only those users that did not have any registered peers at the time of their own activation. This criterion becomes harder and harder to satisfy as the number of registered users rises, and the probability that two Facebook users that registered on the survey application will be connected just by pure chance becomes significant.

Second proxy measure used for endogenous influence is a small community of users who all registered in a short time period, in a timespan of couple of hours in the night of November 27^{th}. For these users it is reasonable to assume that their activation is driven primarily by the endogenous influence. The community itself and their registration dynamics is shown on Figure 2.6. There is a possibility that this is a community of fake Facebook user accounts, activated all from a single source for the purposes of influencing the outcome of the survey. However, this is probably not the case because their actual votes are very heterogeneous - the users in the community are almost equally split between the two survey options presented. The method estimates a sharp rise in endogenously activated users just at the right time while estimate for exogenous influence remains fairly flat.

To further validate the inference method a *randomization strategy* is applied on the referendum2013 social network to see which quantities of interest stay unchanged and

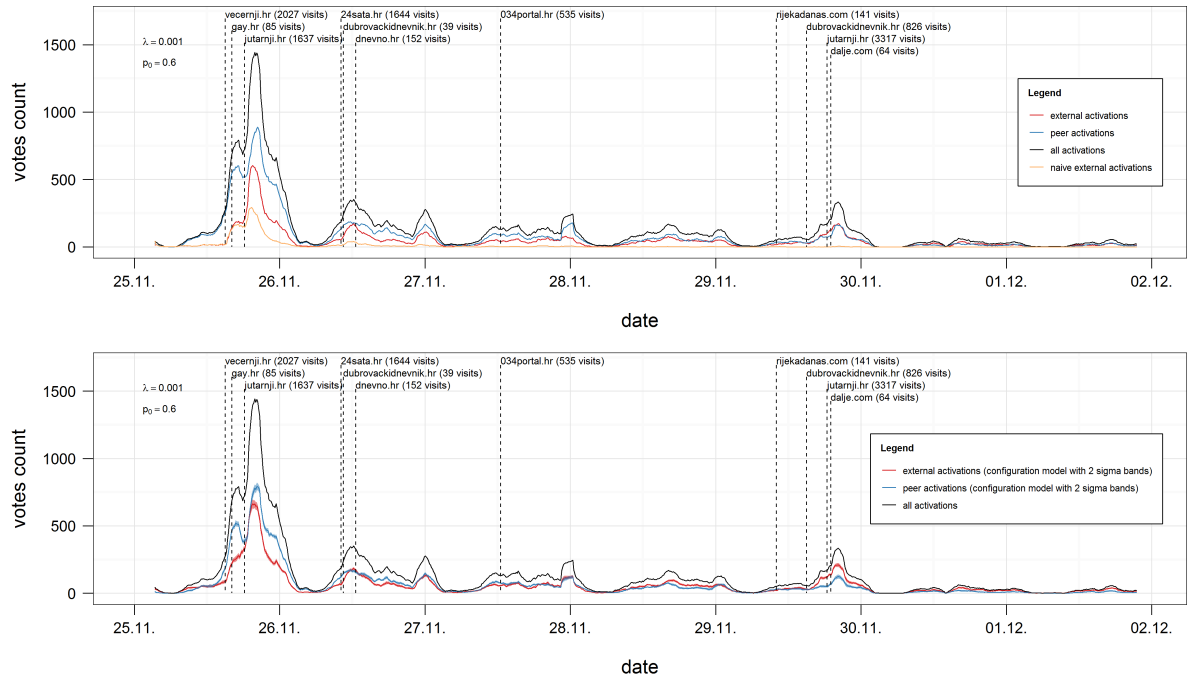


Figure 5.2: Evaluating a direct method of exogenous influence detection on referendum2013 activation cascade. Top of the Figure shows the estimate the absolute number of endogenously and exogenously activated users using Equation 4.2 which assumes exponentially decaying endogenous influence from Equation 3.1. Parameters of endogenous influence which used here ($p_0 = 0.6$ and $\lambda_p = 0.001$) were chosen so that they satisfy an empirical observation of 70% of endogenously activated users during the first day of data collection (Figure 5.1). The sliding window length used for estimation is set to two hours. Bottom Figure shows the same estimates but using a configuration model of social network where friendships connections between users are permuted. This causes the endogenous influence to diminish in magnitude, an effect which is particularly visible during the two periods where endogenous influence is otherwise stronger: 1) first 12 hours, before the publication of the first online news article reporting on the survey application (on vecernji.hr new portal), and 2) few hours during the evening of November 27^{th} where a distinct community of users activated in a short time span.

which are diminished. A particular strategy used is a *configuration model* of a network where the number of friendship connections each user has is kept unchanged but the connections themselves are permuted across the whole network so that users are now randomly connected. This strategy keeps the *global* properties such as degree of nodes in the network intact but disrupts *mesoscale* properties such as communities of users, as well as *local* properties such as local connectivity of users. These latter two are important as they mediate endogenous influence. The rewiring of the friendship connections by the configuration model decouples the activation cascade from the social network, and with it the assumed causal structure underlying the endogenous influence between users. The activation of each user should now be more easily explained by assuming exogenous, rather than endogenous, influence actually took place, which should diminish the estimates of endogenous influence. This is confirmed by repeating the estimation method after rewiring

the friendship connections with the configuration model (Figure 5.2) and observing what happens with the magnitude of endogenous and exogenous influence in the two time periods which are used as a reference for endogenous influence. In the first period, which spans the first 12 hours of user registrations and where the largest fraction of endogenously activated users was originally observed, the magnitude of endogenous influence diminishes and is comparable to the exogenous influence. In the second period, in the few hours on the night of November 27^{th} where strong endogenous influence originating from a single well connected community of users was originally observed, the magnitude of endogenous influence again diminishes.

5.2 Inference of endogenous and exogenous influence on empirical datasets

This section describes the application of the maximum likelihood method for joint inference of endogenous and exogenous influence described in Section 4.3 on all three empirical datasets introduced before - referendum2013 (10175 users), sabor2015 (6909 users) and sabor2016 (3818 users). In addition to the Facebook social networks and users registration cascades which are used for inference, for two of the datasets - sabor2015 and sabor2016, there are also referral links from which users visited the online survey applications. The referral links give us information whether each user followed a link from Facebook, which indicates endogenous influence, or from an external website, which indicates exogenous influence. This allows us to perform evaluation in a more straightforward way, by using evaluation measures for binary classification such as ROC curves and AUC scores. More information on the datasets themselves is available in Section 2.3.

Joint inference is performed by running the Algorithm 4.1 on the three activation cascades. Compared to the inference on simulated activation cascades (Section 4.5) where we operated in discrete time, we now have to perform inference in continuous time. For this the discretization of activation times into 30 minute intervals (the same that are used in user registration histograms in Figures 2.5 and 2.1) is performed. Users that activated in the same 30 minute interval are considered, for all practical purposes, to be activated *at the same time*. Figures 5.3 and 5.4 show the obtained estimates of the absolute number of endogenously and exogenously activated users at each time interval (Equation 4.4) while using EXP and SI as the endogenous influence models respectively.

By using the inferred parameters of endogenous and exogenous influence it is possible to estimate to what extent was each user's activation driven by the one or the other. This is expressed with the measure of *exogenous responsibility* $R^{(i)}(t)$ (Equation 4.5), which is compared with the gold standard labels obtained through referral links in order to cal-

culate ROC curves and AUC scores. The joint inference method is compared with the baseline measure where users that did not have any active peers at the time of their own activation are classified as exogenously activated. Similar as in experiments on simulated activation cascades (Section 4.5) this measure is too conservative and tends to underestimate the true number of exogenously activated peers, especially near the end of the observation period when it becomes increasingly likely that the two users are connected due to chance rather than some underlying endogenous influence between them. A more general version of this baseline method, which is used to calculate the AUC scores, is to use the number of active peers instead of the exogenous responsibility. Figures 5.3 and 5.4 show the results of the inference on empirical Facebook activation cascades. The proposed inference method achieves AUC scores of 0.76 and 0.82 for the sabor2015 and sabor2016 datasets respectively while assuming EXP as the endogenous influence model (AUC_{our} on Figure 5.3). This is significantly better than the baseline method which achieves AUC of 0.68 and 0.78 on the same datasets (AUC_{base} on Figure 5.3). The results are similar if we assume SI as the endogenous influence model instead of EXP - the achieved AUC scores are 0.75 and 0.83 for the proposed inference method (AUC_{our} on Figure 5.4) and 0.68 and 0.78 for the baseline method (AUC_{base} on Figure 5.4) for the sabor2015 and sabor2016 datasets respectively.

Figure 5.5 shows a more clear comparison with the baseline measure in the case where EXP is used as the endogenous influence model. We see that the estimates produced by the proposed inference method for the sabor2015 dataset most closely resemble the number of users that visited the survey application by following other user's Facebook share - the strongest indication of endogenous influence that are obtained from user's referral links. It seems that during the second day of sabor2015 survey the method underestimates the number of endogenously activated users. However, we have to remember that the estimate of endogenously activated users that are obtained from referral links is only an approximation, as users might be activated through other means of indirect communication available in Facebook or even some other social media service - including advertisements, direct messaging or by visiting a Facebook page of the survey application.

A direct benefit of using EXP endogenous influence model instead of SI is that it provides an estimate of the half-decay of the endogenous influence which can be calculated using estimated parameter $\hat{\lambda}$. For the sabor2015 dataset a value of half-decay 10.1 hours is obtained which is consistent with the expectation that the endogenous influence should diminish in the order of a few days. Any endogenous influence beyond this period is more probably sustained by newly activated users rather than users that activated much further in time.

We can see that exogenous influence increases drastically near the end of the obser-

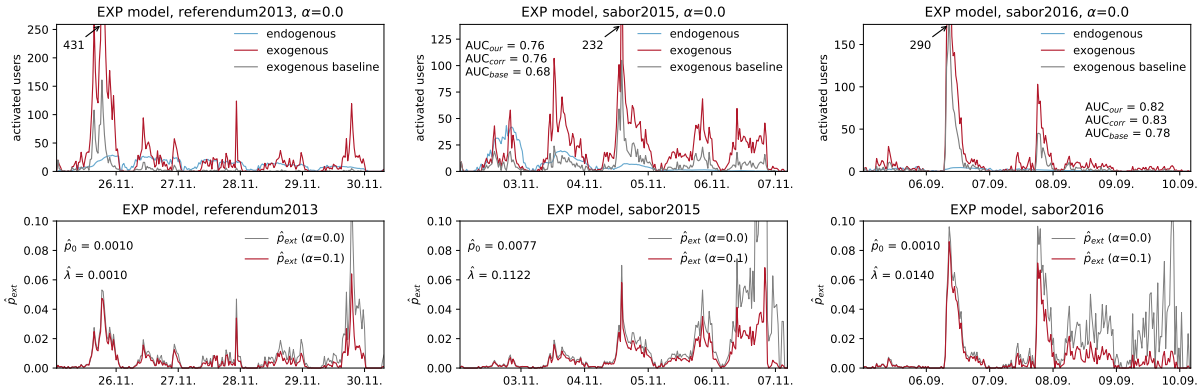


Figure 5.3: Inference of endogenous and exogenous influence on Facebook activation cascades with EXP as the assumed endogenous influence model. The inference is performed on referendum2013 (left), sabor2015 (middle) and sabor2016 (right) datasets in order to estimate endogenous ($\hat{p}_0$,$\hat{\lambda}$) and exogenous $\hat{p}_{ext}(t)$ influence parameters (bottom) as well as absolute number of users activated due to the one or another influence (top). Bottom Figures show the estimated value of exogenous activation probability $\hat{p}_{ext}(t)$ with ($\alpha = 0.1$) and without ($\alpha = 0.0$) correction for the observer bias. The effect of correction is to reduce the overestimate of the exogenous influence near the end of the observation period. AUC scores are calculated to evaluate predictive power of the inference method which uses exogenous responsibility $R^{(i)}(t)$ as a criterion for classifying users into endogenously and exogenously activated, as compared to the proxy obtained using user's referral links (which are only available for sabor2015 and sabor2016 datasets). Inference is performed twice in order to calculate AUC with (AUC_{corr}) and without (AUC_{our}) correction for the observer bias. Estimates are then compared it with the baseline method (AUC_{base}) where, instead of exogenous responsibility, a number of active peers is used as a classification criterion. The achieved AUC scores (AUC_{our}) are 0.76 and 0.82 for sabor2015 and sabor2016 datasets respectivelly, which is higher that AUC scores achieved with the baseline method (AUC_{base}) which are 0.68 and 0.78 for the same datasets. Correction for the observer bias does not influence the predictive power, which is probably due to the fact that the majority of activated users (and with it, the majority of predictive power) is in the first half of the activation cascade where the effect of correction is negligible. Figure 5.5 shows a more detailed comparison with the baseline method.

vation period, a phenomena which seems almost like an anomaly in data. The reason for this the *observer bias* effect which is already mentioned in Section 2.4, and which arises because Facebook Graph API allows to collect only friendship relations between pairs of users which *both* registered on the survey application. The results is that, at the end of the collection period, there is only a subset of friendship connections and so the number of inactive users is underestimated which is crucial for the inference (Equation 3.7). A correction factor is applied in the log-likelihood function to compensate for this bias (Section 4.4), which allows the calculation of more precise estimates of exogenous influence, especially near the end of the observation period where the bias is the most pronounced (Section 5.2). However, because there is less and less users that register as we approach the end of the observation period, the error that this bias introduces in the final estimates is actually not too high. The difference in AUC scores in cases where correction is applied

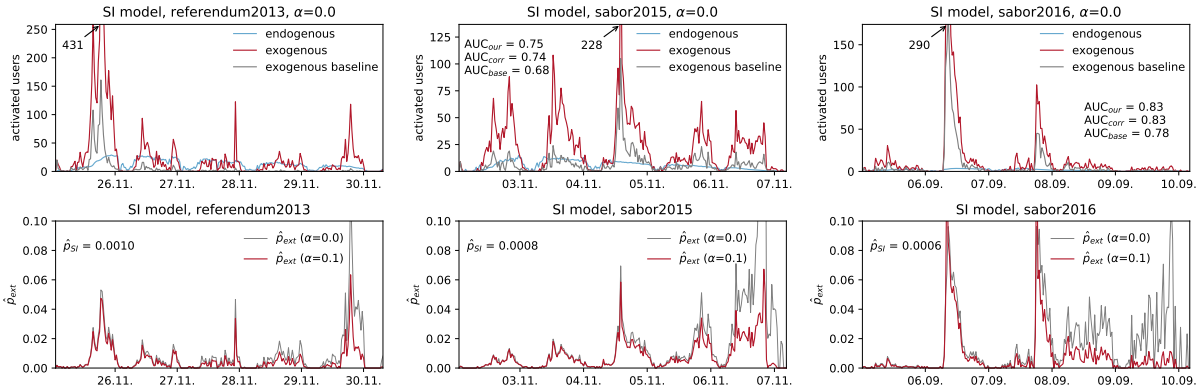


Figure 5.4: Inference of endogenous and exogenous influence on Facebook activation cascades with SI as the assumed endogenous influence model. Here, only one parameter of endogenous influence is inferred - $\hat{p}_{SI}$, otherwise experimental the setup is equivalent to the one in Figure 5.3 where the same inference is performed but assuming EXP endogenous influence model. Again we see a better predictive performance than the baseline method, with AUC scores for the method proposed in this thesis (AUC_{our}) of 0.75 and 0.83 for the sabor2015 and sabor2016 datasets respectively, which is higher than for the baseline method (AUC_{base}) which achieves 0.68 and 0.78 for the same datasets. The effect of applying the correction for the observer bias is similar as for the EXP model - the apparent overestimate of exogenous influence $\hat{p}_{ext}$ near the end of the observation period is corrected, although the contribution to the predictive power as measured by the AUC score is negligible.

($\alpha = 0.1$) or not applied ($\alpha = 0.0$) is in the order of 0.01 points.

Section 4.4 mentions the issue of observer bias. Because of the way the data is collected - only available data is on Facebook users that *eventually* registered on the survey application, there is a tendency to overestimate exogenous influence as we make estimates closer to the end of the data collection (observation) period. This happens because the set of the *observed* inactive users, which is used to estimate the second term of the log-likelihood Equation 3.7, becomes smaller over time, although the true number of inactive users which we did not observe is much larger. Correction for the observer bias is applied through correction factor $c(t)$ (Equation 4.6) in the log-likelihood function (Equation 4.7). The strength of the correction is regulated with the parameter α in the expression for $c(t)$. The interpretation of the correction is that we artificially increase the number of inactive users and in that way obtain a more representative value for the second term in the Equation 4.7 for log-likelihood. Experiments in Figures 5.3 and 5.4 already show the effect of correction for the observer bias by using $\alpha = 0.1$, which is already enough to significantly reduce the overestimation of the exogenous influence near the end of the observation period. Although effect of correction is clearly visible on the inferred parameters of exogenous influence $\hat{p}_{ext}(t)$, its contribution to the estimates of absolute number of activated users is not pronounced. This is because the early periods of activation cascade carry the majority of activated users while their number tends to fall off as we approach the end of the observation period, where the effect of correction is the strongest. To con-

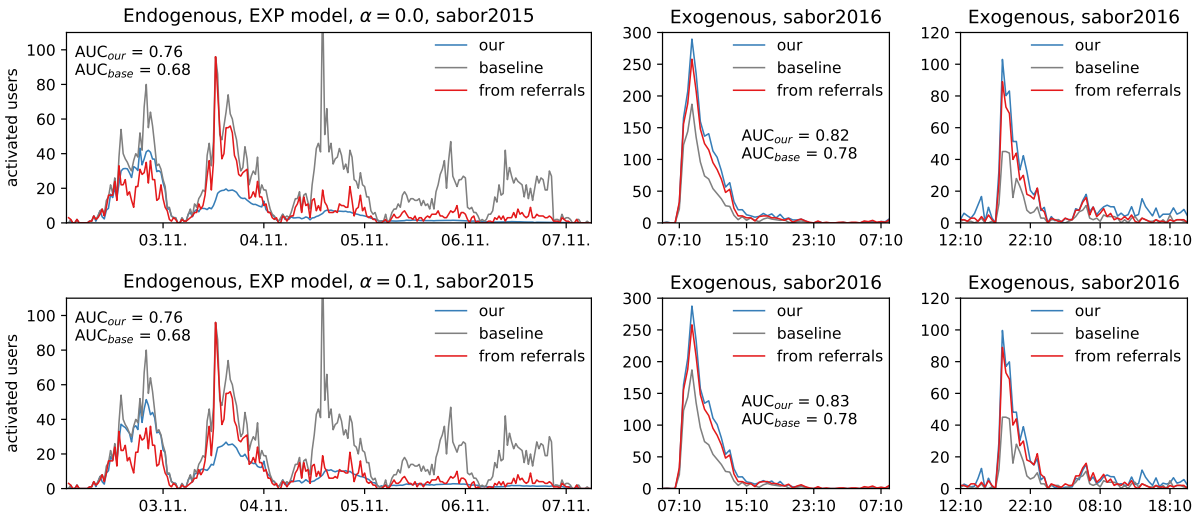


Figure 5.5: Inference of endogenous and exogenous influence on Facebook activation cascades with EXP as the assumed endogenous model - comparison with the baseline method. To aid visual comparison the only endogenously activated users for the sabor2015 dataset and exogenously activated users for the sabor2016 dataset are shown. The estimated number of activated users ("our" for the inference method proposed in this thesis and "base" for the baseline) are compared with the values obtained from user's referral links ("from referrals"). The contribution of applying the correction for the observer bias ($\alpha = 0.1$, bottom) to the predictive power of the inference (AUC score) is negligible as compared to results without the correction ($\alpha = 0.0$, top). The graphs for sabor2015 dataset hint at a possible reason for this. Near the end of the observation period, where the correction is strongest (see, for example, Figure 5.6), there are very little users activated due to someone's else share which is effectively the gold standard for the endogenously activated users which is used to calculate the AUC score.

firm this a more extensive experiments are performed with the wider range of correction factors α up to 0.3 in Figure 5.6. Only sabor2015 and sabor2016 datasets are used in these experiments in order to calculate AUC scores. We see that the effect of correction is the most pronounced near the end of the observation period although the overall predictive performance does not change much. Small corrections ($\alpha = 0.1$) usually lead to small increase in AUC score on the order of 0.01 points, the only exception in Figure 5.6 is the SI model on sabor2015 dataset. Large corrections ($\alpha = 0.3$) usually lead to drop in predictive performance, which is especially pronounced with the EXP model on the sabor2015 dataset. A general conclusion is that the correction for the observer bias should primarily be used as a measure to stabilize the estimates of the parameters of exogenous influence $\hat{p}_{ext}(t)$.

The output of the inference procedure are the parameters of endogenous and exogenous influence from which are used to calculate the value of exogenous responsibility $R^{(i)}(t)$ for each individual user i at the time t of his activation (Equation 4.5. This characterizes each user's activation on the scale from completely endogenous ($R^{(i)}(t) = 0$) to completely exogenous ($R^{(i)}(t) = 1$). It is this measure that is used for calculation of the ROC curves

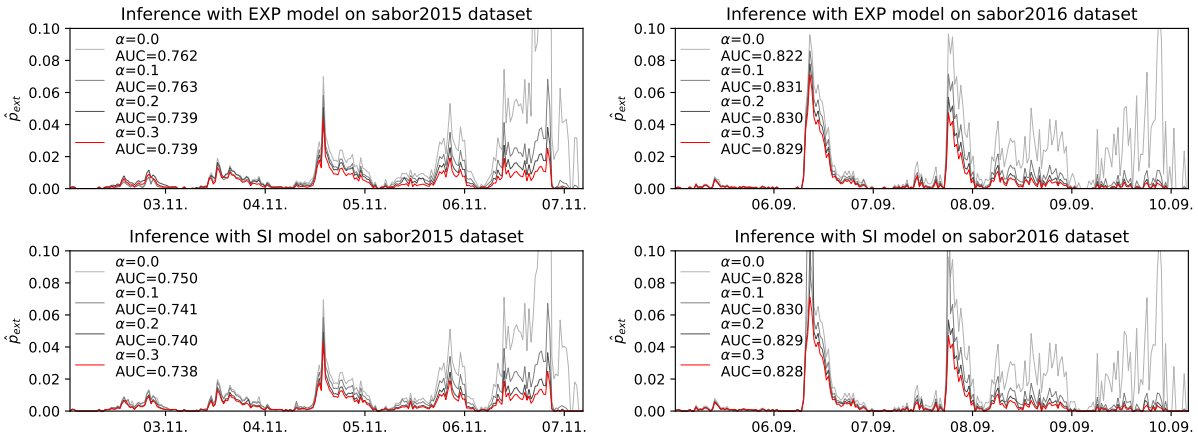


Figure 5.6: Correction for the observer bias while inferring parameters of exogenous influence $\hat{p}_{ext}$ on sabor2015 (left) and sabor2016 (right) datasets. The assumed endogenous influence models are SI (bottom) and EXP (top). The range of the correction factor α is from 0.0 (no correction) to 0.3. The interpretation of the correction is that the set of inactive users is artificially increased as compared to the actually observed value (which we know is an underestimate because of the way the user data is collected). The effect of the correction is the strongest near the end of the observation period where the underestimate of inactive users is the most pronounced. However, the contribution of correction to the overall predictive performance as measured by AUC score is not pronounced - small values of α increase the AUC slightly while large values decrease it below the level of no correction. This is probably because the majority of predictive performance is carried by the activations at early stages of activation cascade. As we approach the end of the activation cascade there is less and less activations (see, for example, Figures 5.4 and 5.4) so the correction is unable to influence the predictive performance much.

and AUC score in all the experiments described in this thesis. Figure 5.7 shows the full distribution of exogenous responsibility across all users in the three datasets. The equivalent histograms on simulated activation cascades were presented in Figures 4.4 and 4.5. As compared to simulated experiments, in experiments on empirical datasets there is no gold standard labels which can be used for evaluation, Instead, a proxy based on user's referral links is used instead. Histograms on Figure 5.7 are divided in three groups based on the type of their referral links - i) unknown and Facebook users without a share, ii) Facebook users with a share and iii) users with a referral link and Facebook users without a share. For referendum2013 dataset there is no information on user's referral links so a distribution across all users is shown. An assumption is that the users that visited the online survey application by following a referral link originating on some external website (for example, an external news media article) should be characterized as being more exogenously influenced. For these users the distribution of exogenous responsibility values should be concentrated near the high values. On the contrary, users that visited the online survey application by following a link originating from Facebook should be characterized as being more endogenously influenced. For these users it is expected that the values of their exogenous responsibility will be more concentrated near the low values,

or at least to be more uniform across the range from 0 to 1. However, not all referral links originating from Facebook should be treated equal, as some are are connected with another user's Facebook share (indicating a direct endogenous influence between the two users) while others are connected with various other indirect forms of Facebook communication, for example news feeds and Facebook pages. These latter ones are sometimes better characterized as being more exogenously influenced, although this cannot be known for sure.

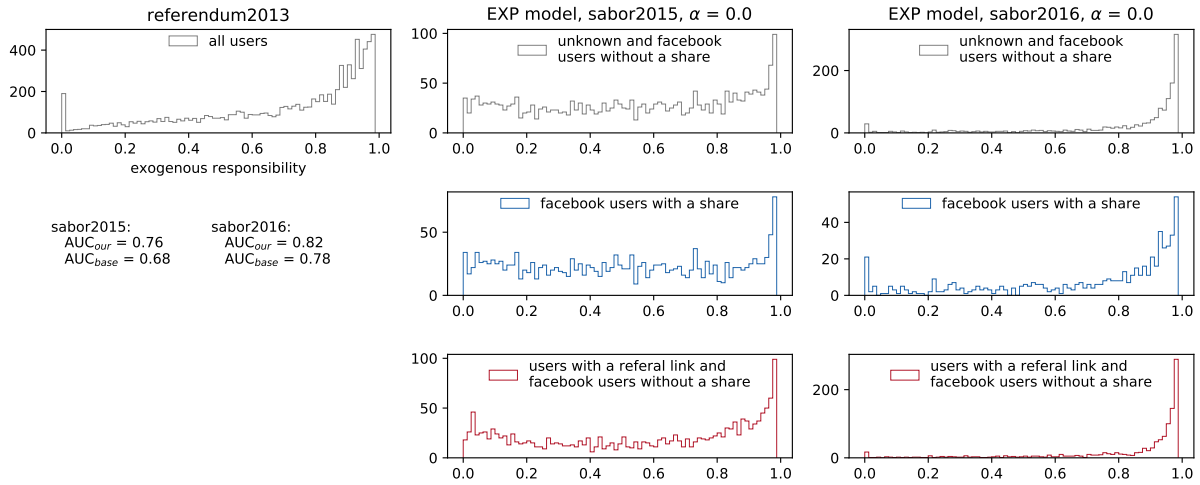


Figure 5.7: Histograms of exogenous responsibilities $R^{(i)}(t)$ for all users which registered in the three online survey applications. The assumed endogenous influence model is EXP and there was no correction for the observer bias ($\alpha = 0.0$). Inference is otherwise equivalent to the experiment in Figure 5.3. Histograms are arranged based on the survey (columns) and the type of user's referral links (rows) which are separated into all users (top), endogenously registered users (middle) and exogenously registered users (bottom). Histograms of exogenously activated users - those whose referral links originate from an external website or from Facebook but that are not linked to a specific share, tend to be more concentrated near the high values of exogenous responsibility. The difference as compared to endogenously activated users is not large, but enough to achieve AUC score (AUC_{our}) of 0.76 and 0.82 for the sabor2015 and sabor2016 datasets respectively. The corresponding AUC scores for the baseline method (AUC_{base}) - where a number of active peers is used as a measure of exogenous influence instead of exogenous responsibility $R^{(i)}(t)$, are lower with values of 0.68 and 0.78 for the sabor2015 and sabor2016 datasets respectively.

5.3 Collective influence in empirical datasets

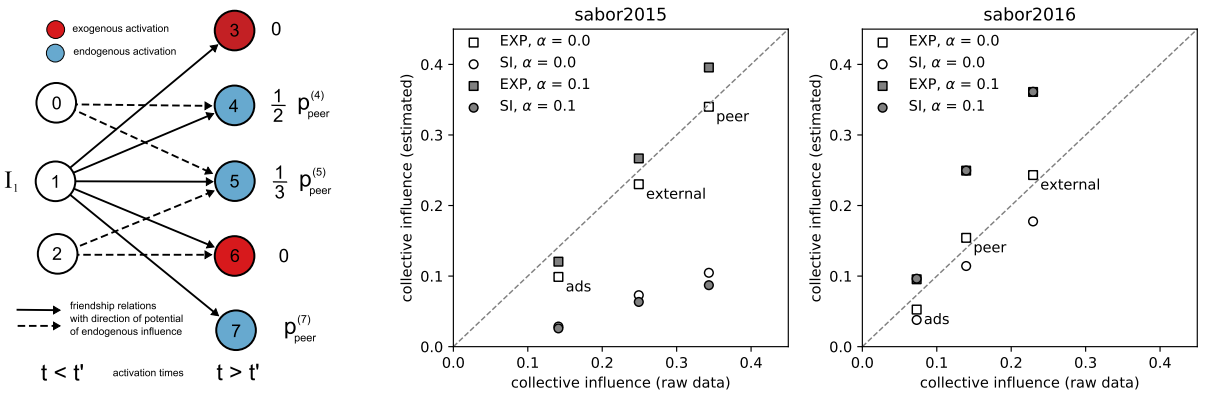
Section 4.7 describes how to calculate *individual influence* of a user (Equations 4.10 and 4.11). It also provides a definition of *collective influence* of a group of users G as the average of individual influences I_i of all users in the group $1/\mathsf{G}\sum_{i\in G} I_i$. Figure 5.8a shows a simple example how to calculate individual influence of a user on a small social network. An underlying requirement is that there is, for each user, an estimate of whether its

activation was due to endogenous or exogenous influence. For this we can use an inference procedure such as the one presented in this thesis or we can try to use a proxy available directly in raw data, for example information on user's referral links. Figure 5.8b shows the collective influence of three groups of users in sabor2015 and sabor2016 Facebook datasets, based on the type of their referral links: i) advertisements (users that followed a link within Facebook advertisement), ii) external (users that followed a link originating from an external website), and iii) peer (users that followed a link from Facebook, for example from another user's share). The collective influence of the three types of users is calculated by using estimates obtained from the inference procedure for both SI and EXP endogenous influence models and with ($\alpha = 0.1$) and without ($\alpha = 0.0$) correction for the observer bias. These estimates of collective influence are compared with the ones calculated using information on user's referral links. Although the magnitude of estimates differ, with the estimates provided by the inference method being typically lower on average on sabor2015 datasets and higher than average on sabor2016 dataset, qualitatively the estimates are proportional to one another. This is especially evident on the sabor2015 dataset while using the EXP as the assumed endogenous influence model. The proportionality of the estimates of collective influence with the ones obtained from the referral links shows that the inference method can reconstruct the underlying information in the referral links which was otherwise *not used* in the inference itself! This shows that the characteristics of the user's activation - whether their activation is more endogenously or exogenously driven, can be inferred from the dynamics of user activation and their mutual social network relationships alone!

Experiment on Figure 5.8b tries to provide an answer to the question which channel of communication - advertisement, endogenous or exogenous, is the most effective in recruiting users with higher collective influence? As there are only two datasets on disposal, the results are not conclusive. In sabor2015 the most influential group are users activated via Facebook itself, which might be due to to the fact that the majority of these users activated fairly recently in the activation cascade (see Figure 2.5). In sabor2016 dataset the most influential group are the users that activated by following a link from an external website, which might be due to the fact that these users also consist a vast majority among all users in the dataset (again, see Figure 2.5).

5.4 Selecting appropriate endogenous influence models

Section 3.3 introduces several endogenous influence models, as well as the model for the exogenous influence. The question remains how to select among them the most



(a) **(b)**

Figure 5.8: Individual and collective influence of users. Figure 5.8a shows a simple example with a small social network neighborhood of eight users $i = \{0,1,2,3,4,5,6,7\}$. An influence I_1 for the user $i = 1$ should somehow summarize the extent to which is he responsible for the endogenous activation $p^{(j)}_{peer}$ of his peers $j = \{3,4,5,6,7\}$ that activated after him ($t_i < t_j$). He can claim responsibility only for the peers $j = \{4,5,7\}$ that activated due to the endogenous influence, but he has to share a part of this claim with users $k = \{0,2\}$ that are also peers of users j and that activated before them ($t_k < t_j$). In this example an assumption is that the partition of influence is equal among users (Equation 4.10) so the total influence of user $i = 1$ is $I_1 = 1/2p^{(4)}_{peer} + 1/3p^{(5)}_{peer} + p^{(7)}_{peer}$. Figure 5.8b shows the collective influence of three groups of Facebook users from the sabor2015 and sabor2016 datasets, based on their referral links - i) advertisements (for users that followed Facebook advertisements that promoted the survey applications), ii) external (for users that followed links from external websites), and iii) peer (for users that followed links from within Facebook). Collective influence is calculated by using estimates of endogenous and exogenous influence from the inference method (y-axis) and by using raw data on referral links (x-axis). In the latter case the values of $p^{(j)}_{peer}$ are from $\{0,1\}$. Facebook advertisements are interesting because they target wide range of users irrespective of their Facebook friendships, which produces similar effect as the exogenous influence.

appropriate influence model. Unlike *parameter estimation*, where the goal is to obtain just a single hypothesis - a set of parameters which completely specifies the model, in *model selection* the goal is to select a family of hypotheses which best describe the data. Typically, model selection is more appropriate if one wants to select a model or a class of models which generalizes best to unobserved data (has the best *predictive performance*) under a variety of circumstances where particular parameters of a model may differ [134]. Also, as each model gives only a partial insight into the underlying phenomena, selecting an appropriate model involves deciding which aspects of the phenomena one wishes to study. In our case, for example, choosing between SI and EXP models is a choice of considering an additional endogenous influence parameter λ which determines the rate of influence decay.

In general, selecting a model with the best predictive performance is not an easy task as the fitness of the model on the *observed* data can mislead us into overestimating its predictive performance on the *unobserved* data. Fitness on the observed data can

always be increased by increasing the *capacity* of the model, although by increasing it too much results in fitting the noise in the data instead of the underlying signal of interest, which reflects poorly on the predictive performance. This *bias/variance* trade off means that overly simple, undetermined, high *bias* models perform poorly on both observed and unobserved data, while overly complex, overdetermined, high *variance* models perform excellent on observed data and poorly on unobserved data. So all model selection methods have to explicitly or implicitly account for the capacity of the model in order to select the models with the best predictive performance.

For example, in Bayesian model selection the model's complexity is implicitly accounted for by using the *marginal likelihood* or *model evidence* as the model selection criteria. Marginalization of the likelihood function over the parameter space has an implicit effect of restricting the complexity of the model beyond what can be supported by the observed data [135]. Because Bayesian approach treats all aspects of modeling as probabilities - models themselves can be interpreted as probability distributions over the space of all possible datasets. Simple models concentrate their probability mass/density to the smaller number of datasets than complex models, but give each of them larger probability. This ensures that complex models will be penalized if data can indeed be explained by a more simple model [136]. Performing a fully Bayesian model selection requires integration of marginal likelihoods over the whole parameter space, which is often infeasible. For this reason there are many other more efficient model selection methods which attempt to approximate capacity of a model in an indirect way.

Maximum likelihood based criteria involves using the maximum value of a likelihood function along with different complexity terms to approximate model evidence. The most commonly used model selection criteria from this category are Bayesian Information Criterion (BIC) [137], Akaike Information Criterion (AIC) [138] and Rissanen's Stochastic Complexity (SC) [139]. All of these measures in general contain two terms - first which is the actual maximum likelihood value and determines goodness-of-fit, and second which approximates model's complexity by a simple expression involving the number of parameters and the number of observations. If the functional forms of the models are the same and they have the same number of parameters their comparison reduces to the generalized likelihood ratio testing [134]. In general the maximum likelihood value is a good approximation to model evidence if the likelihood function itself is sharply peaked over the maximum likelihood parameters, which ensures that the maximum likelihood value and the integral of the likelihood are approximately the same. In the case when this is not satisfied, and the likelihood functions themselves differ significantly in their functional form, the complexity of the model might be better determined by the functional form of the model rather than the number of parameters it contains, which is not captured by

the methods relying on maximum likelihood values.

Out of methods that *explicitly* account for the complexity of models, the two are most prominent - *Minimum Description Length* (MDL) and *Structural risk minimization* (SRM). MDL [45, 134, 140] measures the length in bits of the shortest possible code which describes the data generated by the model. MDL model selection is essentially the same as performing Bayes factor analysis with Jeffrey's prior [134] - a non-informative prior distribution for the parameter space. SRM [141] uses Vapnik-Chervonenkis dimension (VC-dimension) as a measure of model complexity, which is not in the same units as the term for fitness (or "risk") and so their combination is not straightforward [134]. The bounds that VC-dimension provides are very conservative, and can be considered as the *worst-case* estimate of the model's complexity [135].

Two methods that *implicitly* account for the complexity of models are False Discovery Rate (FDR) and Cross-validation (CV). FDR [142] is often used when one wants to select one particular point-hypothesis out of a finite set of hypotheses. It controls the expected proportion of rejected null-hypotheses which were in fact correct ("false discoveries"). In a CV [143] procedure the train and test steps are repeated with a same type of model on multiple disjoint subsets of the observed data in order to select a model that will have good predictive accuracy on unobserved data. In this way model's complexity is incorporated implicitly because models that overfit on the training subset will be penalized by evaluation on the test subset.

In order to justify the choice of exogenous responsibility $R^{(i)}(t)$ (Equation 4.5) as the measure of exogenous influence a comparison of its performance with several other possible measures of influence on empirical datasets is provided. Section 4.5 and Figure 4.8 show a comparison of several different versions of exogenous responsibility (Equations 4.8 and 4.9). Here, a comparison of exogenous responsibility with exogenous activation probability $p^{(i)}_{\text{ext}}$ and endogenous activation probability $p^{(i)}_{\text{peer}}$ is provided. Endogenous activation probabilities are available directly as an output of the inference procedure (Algorithm 4.1). Figure 5.9 shows the ROC curves and the corresponding AUC scores for all the experiments. In order to use ROC curves and AUC score as an evaluation measure the inference is performed only on the two datasets for which there is information on user's referral links - sabor2015 and sabor2016. Experiments also include different combinations of assumed endogenous influence models (SI and EXP) and different corrections for the observer bias - $\alpha = 0.0$ which corresponds to no correction and $\alpha = 0.1$ which corresponds to a slight corrections. These are the same α values that are used in the main experiments shown in Figures 5.3, 5.4 and 5.5. Although neither of the measures performs best under all experimental conditions and datasets, the exogenous responsibility has the most consistent performance across different endogenous influence models and different values of

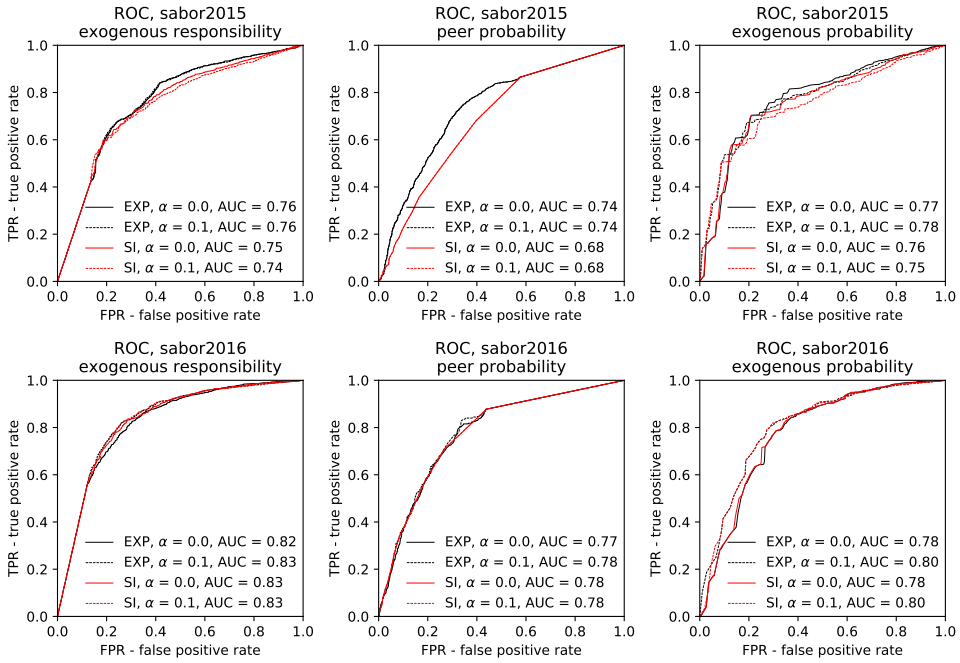


Figure 5.9: Comparison of several possible measures of exogenous influence: (i) exogenous responsibility (left column), (ii) peer probability (middle column), and (iii) exogenous probability (right column). Experiments are performed on two empirical datasets for which the information on user's referral links is available - sabor2015 (top row) and sabor2016 (bottom row), with various combinations of assumed endogenous influence models (SI and EXP) and the correction for observer bias ($\alpha = 0.1$ and $\alpha = 0.0$). Not a single measure performs best in all cases, although exogenous responsibility performs consistently well across different experimental conditions.

correction for the observer bias - being outperformed only by the exogenous activation probability on the sabor2015 dataset.

5.5 Comparison of influence with the structural measures on empirical data

Similarly as in Section 4.8, estimates of individual influence calculated with Equation 4.10, and assuming SI as the endogenous activation model, are compared with several structural measures of influence on all three empirical datasets. Figure 5.10 shows comparison of user's ranks obtained with the influence measure with the four structural influence measures: (i) number of peers which activated *after* a particular user, (ii) number of peers, (iii) activation time, and (iv) number of peers which activated *before* a particular user. Associated Spearman correlation coefficients are calculated to asses the association between these structural measures and the measure of influence proposed in this thesis. Across all datasets, the highest correlation is with the number of peers activated *after* a particular user - with correlations of 0.94, 0.94 and 0.97 for the referendum2013, sabor2015 and sabor 2016 datasets respectively. Somewhat lower are the correlations with

the number of peers - a quantity that does not change over time in the empirical datasets, which are 0.67, 0.75 and 0.78 for the referendum2013, sabor2015 and sabor2016 datasets respectively. These results align with our intuition, as it is expected that users with more peers have more opportunities to spread their influence, and for the influence to act *forwards* in time rather than *backwards*. This is supported by the correlations with the number of peers activated *before* a particular user which are much lower - 0.17, 0.42 and 0.50 for the referendum2013, sabor2015 and sabor2016 datasets respectively, and correlations with the activation time of a particular user which are *negative* - -0.57, -0.64 and -0.32 for the same datasets respectively! These results are qualitatively equivalent to the ones obtained on simulated data (Figure 4.11).

Figure 5.11 shows the comparison on the same type of scatter plots but instead of the before mentioned simple structural measures five baseline structural measures commonly used in literature are used: (i) degree centrality (which is identical to the number of peers), (ii) Pagerank centrality [144], (iii) eigenvector centrality [145], (iv) hubs centrality [146, 147], and (v) authorities centrality [146, 147]. Again, Spearman correlation coefficient is taken as the measure of correspondence with the influence calculated with Equation 4.10. Overall, the strongest correlation is with the Pagerank centrality - 0.86, 0.90 and 0.92 for the referendum2013, sabor2015 and sabor2016 datasets respectively, while the weakest is with the hubs centrality - 0.32, 0.50 and 0.55 for the same datasets respectively. Relatively high correlation with all above mentioned structural measures indicates that there is at least some consensus on which users are overall more or less influential.

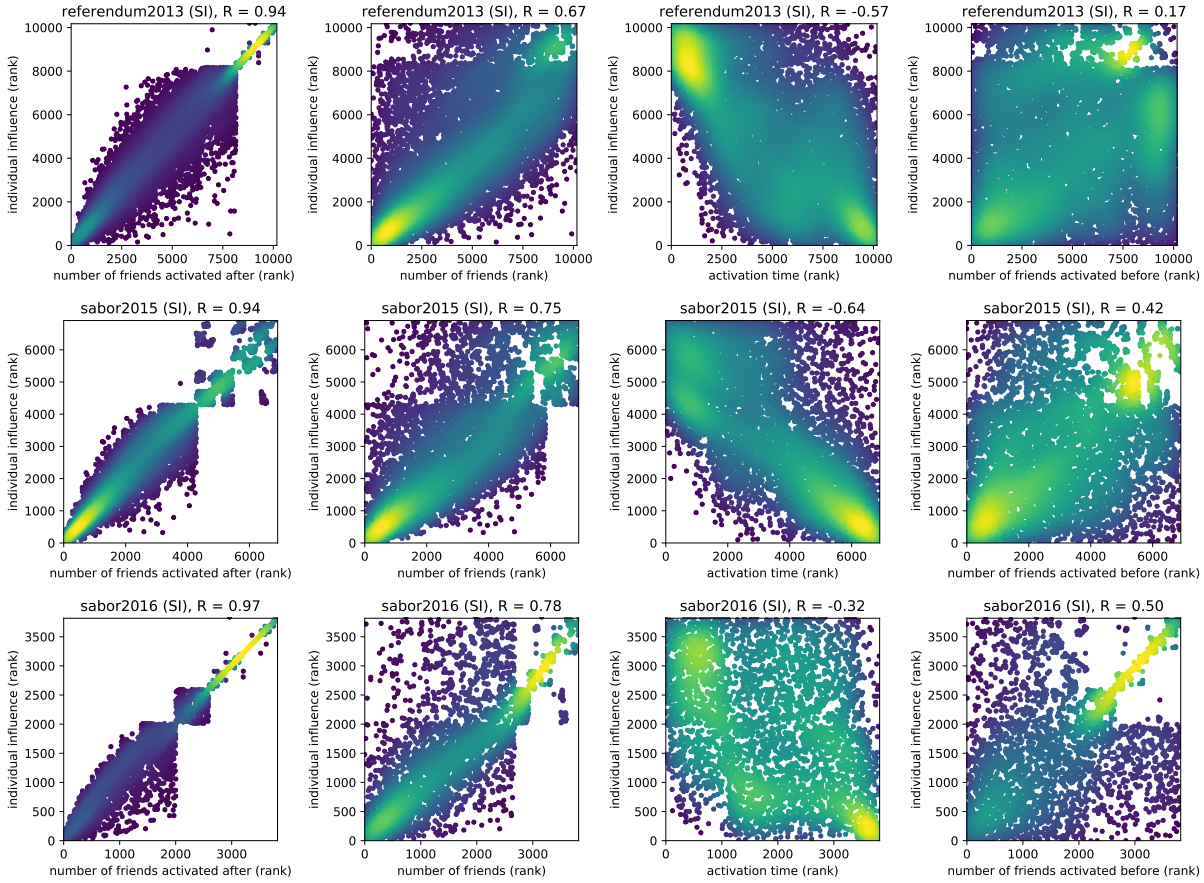


Figure 5.10: Comparison of user influence with four simple structural measures on empirical datasets. Scatter plots show comparisons of ranks of all users obtained by user influence calculated with equation 4.10, with SI as the assumed endogenous influence model, and four simple structural measures of influence. Spearman correlation coefficient is calculated as the measure of association, and points on the scatter plots are colored based on local density so that areas of higher density are colored yellow while areas of lower density are colored blue. The highest correlation across all datasets is with the number of peers activated *after* a particular user, followed by the correlation with the number of peers, which is intuitive as it is expected that more peers and early activation provides more opportunities to spread the influence. On the other hand, influence is less correlated with the number of peers that activated *before* a particular user, and *negatively* correlated with the activation time, which is unsurprising because it is expected that the influence between users propagates *forwards* rather than *backwards* in time.

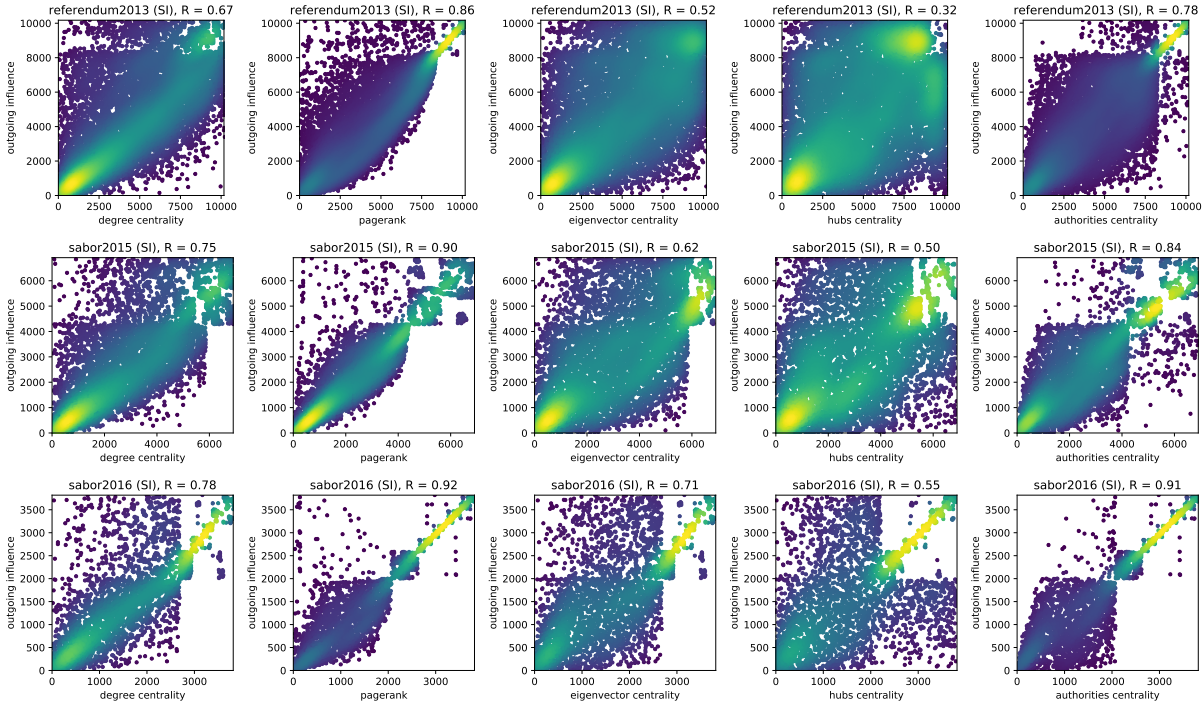


Figure 5.11: Comparison of user influence with five baseline structural measures on empirical datasets. Similar as in Figures 4.11 and 5.10, scatter plots show ranks obtained by the measure of influence calculated with Equation 4.10, with SI as the assumed endogenous influence model, and ranks obtained by five baseline structural measures. The Spearman coefficient of correlation is used as the measure of correspondence between the ranks. The strongest correlation is achieved with the Pagerank and the lowest with the hubs centrality.

Chapter 6

Conclusion

Web-based news media, online social networks and online information services have an immense influence on the way how people interact with the world around them and between themselves. Online social networks are less than 15 years old and already made tectonic changes in the information and marketing industry. Out of the 10 most valuable world companies by market capitalization*, five of them - Microsoft, Apple, Amazon, Alphabet (parent company of Google) and Facebook, are primarily information companies and create most of their value, directly or indirectly, by monitoring activity of the people using their services and understanding patterns of influence between them.

However, user's increased awareness of the importance of data privacy and the potential for manipulation by third parties calls for an increased research effort into quantifying to what extent digital footprints reveal about individuals. For example, it was shown that personality traits of users could be estimated indirectly from the content with which they interact [62] and that users could unknowingly take part in misinformation (fake news) spreading [53]. The research presented in this thesis shows that much can be learned about users engagement even by just observing a friendship network between registered users and their registration times - effectively just a single activation cascade. This data alone is sufficient to estimate whether users were predominantly influenced by their peers, which corresponds to endogenous influence, or by the external factors such as mass media, which corresponds to exogenous influence. Both exogenous [18] and endogenous [59] factors are known to have a significant impact on user's activity.

Method developed as a part of this doctoral research can be used for inferring endogenous and exogenous influence between users of an online social network. The only information needed for inference are the social connections between users and a single activation cascade between them. The hypothesis is that these two influences differ in their statistical properties and that this can be exploited to disentangle their relative

*Data on market capitalization retrieved from `https://ycharts.com`.

magnitude in empirical data. An underlying assumption behind the method is that the endogenous influence propagates between users and so is somehow dependant on the underlying social network structure, while exogenous influence, being external to the social network itself, is independent of its structure. The first approach was to devise a simple method which uses a statistical threshold to estimate the number of exogenously activated users [101]. The second approach was to devise a fully probabilistic method which incorporates both influences jointly and uses a maximum likelihood method to infer the parameters of influence. Using these we can characterize the activation of each user as being more endogenously or exogenously driven, which allows us to estimate individual influence of each user towards all of his peers while correcting for the confounding exogenous influence, even without knowing exactly who influenced whom. This information could, in principle, be used for reconstructing the most probable deterministic influence path, even though some parts of the pathway are inherently unobservable and might be attributed to factors outside of the social network itself.

The proposed method is flexible enough to incorporate additional information regarding the activation cascade and any characteristics of the users or the social network itself. A starting point could be features which are included in the unified model of social influence [148], which is also a likelihood-based model. The assessment of the computational scalability of the method is also performed, concentrating on the scalability of inference rather than scalability of modeling as is more commonly done [149]. Inference is performed through a maximum likelihood method that utilizes efficient numerical solvers, allowing the method to scale on social networks of over 10000 users.

Question remains as to the applicability of our inference method outside of the use cases described in this thesis. The method can be viewed as a part of a larger framework which aims to efficiently characterize the types of influence in information spreading. This framework could be used, for example, to elucidate the role of external factors in misinformation spreading over online social networks [150]. Outside of the domain of social network, the inference of endogenous and exogenous influence could be used in a wider context of dynamical systems modeling [50].

The empirical analysis is performed on data of over 20 thousand Facebook users obtained through three unique Facebook political survey applications. The methodology allows us to estimate, for the first time, the most probable source of influence for each active individual in the survey, and assess the overall influence of different media channels for spreading of the information (peer communication, Facebook advertisements, or external news media) using only a single activation cascade. Besides inference methodology the thesis also discusses valuable guidelines for researchers interested in collecting their own online social network data in an ethically principled way, while at the same

time satisfying requirements for reproducible research. The source codes of the survey applications are freely available on a public open source code repository [†].

The most challenging part of any future work in modeling and inference of influence in online social networks is the data collection and management part, along with the technical, methodological and ethical issues which came along with it. Unlike traditional survey methods, where data is manually entered and the researcher usually has a complete oversight of the data collection process [151], online social networks provide an opportunity to collect much larger amount of data on user activity. However, many of the standard practices for social science research have to be adjusted. Most notable example is a requirement informed consent - a requirement that users are adequately informed on the data collection process and gave an explicit permission for their data to be used in research. Considering that user's data is often collected automatically by the online service provider, usually under a very broad terms of use agreement aimed primarily for marketing research, question remains to what extent can this data be used in academic research. This is why most academic studies using online social networks data are observational in nature and seldom satisfy a requirement of informed consent for all of the users [59], which often raises ethical concerns [2, 75]. Performing a study where explicit informed consent is mandatory heavily restricts the number of users willing to participate, even when researchers are working internally within the online service provider and are in position to seek informed consent from large number of users automatically. A notable example is a study [61] on Facebook where survey was presented to around 1.3 million users, which in the end managed to collect only 7730 responses.

For external researchers data collection on a social network service entails using an using official API or a third-party application. Without an easy automatic access to all users of a service recruitment usually proceeds organically from user to user, mimicking a form of snowball sampling. In this way it is the most eager users that are recruited first, which is in fact crucial for mobilizing the less motivated and weakly connected users although it biases our sample of users. The effect that highly influential users have on mobilization might easily dominate the one from mass media [152]. However, major publicized events such as elections and referendums serve as potent catalyzers for mobilizing users - a fact that we exploited by using online political survey applications for collecting user data. Survey applications were hosted on a separate web pages and used Facebook Graph API for authorization of users, which allowed us to collect activation cascades and friendship connections of over 20 thousand users in total. Considering that collecting data through online social networks is only possible for a decade or so,

[†] `https://github.com/devArena/referendum2013.hr`
`https://bitbucket.org/marin/sabor2015.hr`

standards and practices are still forming under the constant pressure of technological changes. Regardless, we tried to follow current recommended ethical practices [14, 15].

There are several limitations to the methodology which indicate potential directions for future research. First, the way endogenous influence is modeled can be greatly improved. The methodology currently requires a predefined closed form of endogenous influence whose choice implies a particular microscopic influence interaction between users. In this thesis it is showed how one can choose between several competing models of endogenous influence by evaluating their predictive power on the empirical activation cascades. More sophisticated forms of model selection are possible, some of which are mentioned in Section 5.4. Ideally, we would want to have a non-parametric model of endogenous influence whose capacity is automatically adjusted with the observed data so as to prevent overfitting. Second, although the model of influence allows this, the possibility of assigning different propensities for endogenous and exogenous influence to the users is not explored. This could be done, for example, by dividing users into groups and inferring separate models of endogenous and exogenous influence for each group of users, or by including additional user covariates into the influence model. Covariates could be derived from demographic variables which are usually available in online social network datasets, or from information on the activity of users such as their interaction with different content. However, inferring a more expressive influence models will introduce more uncertainty into the estimates, or could even prove to be unfeasible without imposing additional constraints in the inference method, considering that there is only a single activation cascade [153, 154]. This should not be a problem for use cases where multiple activation cascades are available. Adding covariates for the users should also have an additional benefit, as there is a chance to correct for the potential confounding effects arising from the observed characteristics of users, allowing us to disentangle effects of *homophily* from the true social influence. For example, it is expected that users that share political orientation respond differently to each others influences, as compared to the users that do not.

The main contributions of this thesis are the definition of the model of exogenous and endogenous influence in social networks, an inference method which uses a single activation cascade to infer parameters of this model from empirical data, and an extensive evaluation of this inference method on both simulated and empirical data consisting of over 20 thousand Facebook users which participated on several online political survey applications.

Appendix A

Code and data availability

Due to the Facebook's Platform Policy * we are not allowed to publicly release any Facebook-derived data, including personal information and friendship relations between our users. Friendship networks and registration times needed to reproduce the results of this paper are available upon a reasonable request and only after signing the following Data Access Agreement †:

> Upon receiving this dataset you agree to following terms: (i) You will only use the dataset for the purpose of reproducing and validating the results of our study; (ii) You will not attempt to deanonymize the dataset or in any other way compromise the identity or privacy of users contained in it; and (iii) You will not further share, distribute, publish, or otherwise disseminate the dataset.

Source code of the Facebook online survey applications through which we collected referendum2013 and sabor2015 datasets are available on public Github repositories ‡ §.

*https://developers.facebook.com/policy
†https://goo.gl/forms/IxINFkeBSJpDuzRv2
‡https://github.com/devArena/referendum2013.hr
§https://bitbucket.org/marin/sabor2015.hr

Appendix B

Terms of use and privacy policy of the Facebook applications

Disclaimer that we used on front web pages of referendum2013.hr and sabor2015.hr online survey applications was placed next to the registration button and stated which Facebook data we are collecting from users, how will the data be used and how will it be visible to other users of the survey. A full text (in Croatian) of the disclaimer * is the following:

> "U svrhu ovoga istraživanja prikupljamo podatke o vašem stavu o referendumskom pitanju, kao i određene podatke s Facebooka (Facebook identifikacijski broj, godinu rođenja, lokaciju, spol i popis vaših prijatelja), čime želimo dobiti uvid u načine kako međusobna poznanstva utječu na stavove korisnika Facebooka. Vaš individualni odgovor na referendumsko pitanje i podaci o vašem profilu neće biti vidljivi drugim korisnicima, već će samo biti vidljiv anonimni prosjek odgovora. Jedino ćete vi vidjeti prosjek odgovora vaših prijatelja. Istraživači garantiraju da prikupljeni podaci neće biti korišteni ni u koje druge svrhe osim znanstveno-istraživačke. Prijavom na ovaj upitnik potvrđujete da ste suglasni s ovim pravilima korištenja."

In addition to this, the web pages of our online survey application also provided separate web pages with the Frequently Asked Questions (FAQ) †, terms of use ‡ and privacy policy. The link to the privacy policy was provided to the users upon authorization with their Facebook credentials through official Facebook API interface. The full text (in

*https://github.com/devArena/referendum2013.hr/blob/master/referendum/templates/logged_out.html
https://bitbucket.org/marin/sabor2015.hr/src/master/sabor2015/sabor2015/templates/disclaimer.j2

†https://github.com/devArena/referendum2013.hr/blob/master/static/pitanja.html//https://bitbucket.org/marin/sabor2015.hr/src/master/sabor2015/sabor2015/templates/faq.j2

‡https://bitbucket.org/marin/sabor2015.hr/src/master/sabor2015/sabor2015/templates/uvjeti_koristenja.j2

Croatian) of the privacy policy [§] for the sabor2015.hr online survey application is the following:

> Pravila privatnosti
>
> 1. Web servis sabor2015.hr (u daljnjem tekstu Servis) koristi kolačiće za pružanje boljeg korisničkog iskustva. Nastavkom korištenja Servisa slažete se s korištenjem kolačića.
>
> 2. Servis će za svakog korisnika uz prethodno dopuštenje pohraniti:
>
> 2.1. Odgovori korisnika na anketu o izborima
>
> izborna jedinica izborna lista stranaka za koju će korisnik glasati na izborima lista stranaka koje korisnik simpatizira postotak za koji korisnik očekuje da će stranka koju podržava osvojiti na izbora vrijeme glasovanja
>
> 2.2. Facebook podaci
>
> Facebook identifikacijski broj, godina rođenja, lokacija, spol, popis anonimiziranih identifikatora prijatelja koji već koriste naš Servis, preusmjeravajući link – link s koje stranice je došao korisnik pri registraciji
>
> Prikupljanje Facebook podataka ide isključivo preko službenih Facebook sučelja i u skladu je sa svim pravilima za zaštitu privatnosti korisnika: https://developers.facebook.com/policy/.
>
> Anonimizirani Facebook identifikacijski broj jedinstven je za pojedinog korisnika i ovaj Servis te se na osnovu njega ne može doći do stvarnog Facebook računa korisnika. Ne pohranjujemo podatke vezane uz račun korisnika poput email adrese ili imena i prezimena.
>
> 3. Vidljivost i korištenje podataka
>
> 3.1. Vaši pojedinačni odgovori na anketu o izborima (2.1.) i podaci s vašeg Facebook profila (2.2) neće biti vidljivi drugim korisnicima Servisa. 3.2. Vaši prijatelji koji koriste Servis mogu vidjeti samo sumarne glasove svojih Facebook prijatelja (2.1), i to samo onih koji su također korisnici Servisa. Nijedan korisnik Servisa nemaju informaciju kako je i tko od njegovih prijatelja glasovao. 3.3. Sumarna statistika glasova (2.1.) preko svih korisnika Servisa može postati javno dostupna. 3.4. Istraživači će koristiti anonimizirane i sumarne podatke (2.1.) i (2.2.) za znanstvene istraživanja, te će kao takvi biti dostupni znanstvenoj zajednici.
>
> 4. Sva komunikacija s našim Servisom, bazom podataka i Facebook serverom ide preko kriptirane sigurne veze.

[§] `https://bitbucket.org/marin/sabor2015.hr/src/master/sabor2015/sabor2015/templates/pravila_privatnosti.j2`

Appendix C

Implementation details of inference methodology

In order to make our statistical inference as efficient as possible we vectorize all numerical operations using array and matrix primitives within Numpy and Scipy libraries. We store the adjacency matrix of our Facebook friendship network as Scipy's compressed sparse column (CSC) matrix* which allows us to use efficient vectorized implementations of matrix addition and multiplication as well as fast matrix-vector products. A common operation is a selection of all friendship connections towards peers that activated either *before* or *after* a particular user or group of users. This is needed in many places, for example in calculating the endogenous influence in Equation 3.1, expressions for SI model in Equation 3.2 and EXP model in Equation 3.3, expression for the likelihood in Equation 3.6 and log-likelihood with (Equation 4.7) and without (Equation 3.7) the correction for the observer bias, expression for the individual influence in Equations 4.10 and 4.11, and expression for the absolute number of users activated due to endogenous or exogenous influence in Equation 4.4. Here we exploit the fact that these selections can be efficiently performed by sorting the adjacency matrix by the activation time of users and selecting users that activated within a particular time period using *range operator*. This gives performance benefits as slicing a predefined range of a matrix is more efficient than random indexing. This is illustrated in Figure C.1.

*https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csc_matrix.html

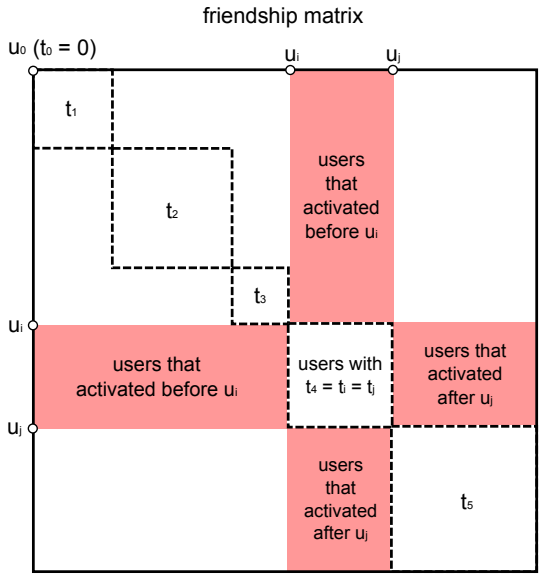


Figure C.1: Reordering of friendship matrix so that users are arranged by their activation times. This allows us to select all peers that activated before or after the users with efficient range operator instead of logical indexing.

Bibliography

[1] Borge-Holthoefer, J., Rivero, A., García, I., Cauhé, E., Ferrer, A., Ferrer, D., Francos, D., Iñiguez, D., Pérez, M., Ruiz, G., Sanz, F., Serrano, F., Viñas, C., Tarancón, A., Moreno, Y., "Structural and dynamical patterns on online social networks: The spanish may 15th movement as a case study", PLoS ONE, Vol. 6, No. 8, 2011.

[2] Kramer, A., Guillory, J., Hancock, J., "Experimental evidence of massive-scale emotional contagion through social networks", Proceedings of the National Academy of Sciences of the United States of America, Vol. 111, No. 24, 2014, pp. 8788-8790.

[3] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N., "Tastes, ties, and time: A new social network dataset using facebook.com", Social Networks, Vol. 30, No. 4, 2008, pp. 330-342.

[4] Karsai, M., Iñiguez, G., Kaski, K., Kertész, J., "Complex contagion process in spreading of online innovation", Journal of The Royal Society Interface, Vol. 11, No. 101, 2014.

[5] Guille, A., Hacid, H., Favre, C., Zighed, D., "Information diffusion in online social networks: A survey", SIGMOD Record, Vol. 42, No. 2, 2013, pp. 17-28.

[6] De Domenico, M., Lima, A., Mougel, P., Musolesi, M., "The anatomy of a scientific rumor", Scientific Reports, Vol. 3, 2013.

[7] Najar, A., Denoyer, L., Gallinari, P., "Predicting information diffusion on social networks with partial knowledge", in Proceedings of the 21st Annual Conference on World Wide Web Companion, ser. WWW '12, 2012, pp. 1197-1203.

[8] Yang, J., Leskovec, J., "Modeling information diffusion in implicit networks", in Proceedings - IEEE International Conference on Data Mining, ser. ICDM '10, 2010, pp. 599-608.

[9] Adar, E., Adamic, L., "Tracking information epidemics in blogspace", in Proceedings - 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2005, Vol. 2005, 2005, pp. 207-214.

[10] Cha, M., Mislove, A., Gummadi, K., “A measurement-driven analysis of information propagation in the flickr social network”, in WWW’09 - Proceedings of the 18th International World Wide Web Conference, 2009, pp. 721-730.

[11] Kwak, H., Lee, C., Park, H., Moon, S., “What is twitter, a social network or a news media?”, in Proceedings of the 19th International Conference on World Wide Web, WWW ’10, 2010, pp. 591-600.

[12] Bakshy, E., Rosenn, I., Marlow, C., Adamic, L., “The role of social networks in information diffusion”, in WWW’12 - Proceedings of the 21st Annual Conference on World Wide Web, 2012, pp. 519-528.

[13] Hu, Y., Manikonda, L., Kambhampati, S., “What we instagram: A first analysis of instagram photo content and user types”, 2014, pp. 595-598.

[14] Salganik, M., Bit by Bit: Social Research in the Digital Age. Princeton University Press, 2017.

[15] Kosinski, M., Matz, S., Gosling, S., Popov, V., Stillwell, D., “Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines”, American Psychologist, Vol. 70, No. 6, 2015, pp. 543-556.

[16] Pan, W., Dong, W., Cebrian, M., Kim, T., Fowler, J., Pentland, A., “Modeling dynamical influence in human interaction”, IEEE Signal Processing Magazine, Vol. 29, No. 2, 2012, pp. 77-86.

[17] Basu, S., Choudhury, T., Clarkson, B., “Learning human interactions with the influence model”, MIT Media Laboratory Vision and Modeling Technical Report, Tech. Rep., 2001, Available: https://alumni.media.mit.edu/~sbasu/papers/InfluenceModel_TR_539.pdf

[18] Aral, S., Nicolaides, C., “Exercise contagion in a global social network”, Nature Communications, Vol. 8, 2017.

[19] Granovetter, M., “Threshold Models of Collective Behavior”, American Journal of Sociology, Vol. 83, No. 6, 1978, pp. 1420.

[20] Watts, D., “A simple model of global cascades on random networks”, Proceedings of the National Academy of Sciences of the United States of America, Vol. 99, No. 9, 2002, pp. 5766-5771.

[21] Snijders, T., Koskinen, J., Schweinberger, M., “Maximum likelihood estimation for social network dynamics”, Annals of Applied Statistics, Vol. 4, No. 2, 2010, pp. 567-588.

[22] Barrat, A., Barthélemy, M., Vespignani, A., Dynamical processes on complex networks. Cambridge University Press, 2008.

[23] Newman, M., Networks: An Introduction. Oxford University Press, 2010.

[24] Kempe, D., Kleinberg, J., Tardos, E., “Maximizing the spread of influence through a social network”, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 137-146.

[25] Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., Kellerer, W., “Outtweeting the twitterers - predicting information cascades in microblogs”, in Proceedings of the 3rd Wonference on Online Social Networks, ser. WOSN’10, 2010, pp. 3–3.

[26] Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., Motoda, H., “Learning diffusion probability based on node attributes in social networks”, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 6804 LNAI, 2011, pp. 153-162.

[27] Leskovec, J., Adamic, L., Huberman, B., “The dynamics of viral marketing”, ACM Transactions on the Web, Vol. 1, No. 1, 2007.

[28] Centola, D., “The spread of behavior in an online social network experiment”, Science, Vol. 329, No. 5996, 2010, pp. 1194-1197.

[29] Pastor-Satorras, R., Vespignani, A., “Epidemic spreading in scale-free networks”, Physical Review Letters, Vol. 86, No. 14, 2001, pp. 3200-3203.

[30] Aral, S., Muchnik, L., Sundararajan, A., “Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks”, Proceedings of the National Academy of Sciences of the United States of America, Vol. 106, No. 51, 2009, pp. 21 544-21 549.

[31] Shalizi, C., Thomas, A., “Homophily and contagion are generically confounded in observational social network studies”, Sociological Methods and Research, Vol. 40, No. 2, 2011, pp. 211-239.

[32] Anagnostopoulos, A., Kumar, R., Mahdian, M., “Influence and correlation in social networks”, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 7-15.

[33] Quattrociocchi, W., Caldarelli, G., Scala, A., "Opinion dynamics on interacting networks: Media competition and social influence", Scientific Reports, Vol. 4, 2014.

[34] Lu, X., Brelsford, C., "Network structure and community evolution on twitter: Human behavior change in response to the 2011 japanese earthquake and tsunami", Scientific Reports, Vol. 4, 2014.

[35] Myers, S., Zhu, C., Leskovec, J., "Information diffusion and external influence in networks", in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 33-41.

[36] Anagnostopoulos, A., Brova, G., Terzi, E., "Peer and authority pressure in information-propagation models", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 6911 LNAI, No. PART 1, 2011, pp. 76-91.

[37] Tang, J., Sun, J., Wang, C., Yang, Z., "Social influence analysis in large-scale networks", in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 807-815.

[38] Saito, K., Nakano, R., Kimura, M., "Prediction of information diffusion probabilities for independent cascade model", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 5179 LNAI, No. PART 3, 2008, pp. 67-75.

[39] Steeg, G., Galstyan, A., "Information transfer in social media", in WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web, 2012, pp. 509-518.

[40] Srivastava, A., Chelmis, C., Prasanna, V., "Influence in social networks: A unified model?", in ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2014, pp. 451-454.

[41] Guilbeault, D., Becker, J., Centola, D., "Complex contagions: A decade in review", in Complex Spreading Phenomena in Social Systems, Sune Lehmann, Y.-Y. A., (ur.). Springer International Publishing, 2018, pp. 3-25.

[42] Chelmis, C., Prasanna, V., "The role of organization hierarchy in technology adoption at the workplace", in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013, 2013, pp. 8-15.

[43] Saito, K., Kimura, M., Ohara, K., Motoda, H., "Selecting information diffusion models over social networks for behavioral analysis", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 6323 LNAI, No. PART 3, 2010, pp. 180-195.

[44] Saito, K., Kimura, M., Ohara, K., Motoda, H., "Detecting changes in information diffusion patterns over social networks", ACM Transactions on Intelligent Systems and Technology, Vol. 4, No. 3, 2013.

[45] Grunwald, P., "Model selection based on minimum description length", Journal of Mathematical Psychology, Vol. 44, No. 1, 2000, pp. 133-152.

[46] Burnham, K. P., Anderson, D. R., Model Selection and Multimodel Inference: A Practical Information-theoretic Approach. Springer-Verlag, 2002.

[47] Machta, B. B., Chachra, R., Transtrum, M. K., Sethna, J. P., "Parameter space compression underlies emergent theories and predictive models", Science, Vol. 342, No. 6158, oct 2013, pp. 604–607.

[48] Shalizi, C. R., "Advanced data analysis from an elementary point of view", https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf, accessed: 2019-05-12.

[49] Onnela, J.-P., Reed-Tsochas, F., "Spontaneous emergence of social influence in online systems", Proceedings of the National Academy of Sciences of the United States of America, Vol. 107, No. 43, 2010, pp. 18 375-18 380.

[50] Argollo De Menezes, M., Barabási, A.-L., "Separating internal and external dynamics of complex systems", Physical Review Letters, Vol. 93, No. 6, 2004, pp. 068 701-1-068 701-4.

[51] Karsai, M., Iniguez, G., Kikas, R., Kaski, K., Kertész, J., "Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading", Scientific Reports, Vol. 6, 2016.

[52] González-Bailón, S., Borge-Holthoefer, J., Rivero, A., Moreno, Y., "The dynamics of protest recruitment through an online network", Scientific Reports, Vol. 1, 2011.

[53] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., Zittrain, J. L., "The science of fake news", Science, Vol. 359, No. 6380, 2018, pp. 1094–1096.

[54] Brach, P., Epasto, A., Panconesi, A., Sankowski, P., "Spreading rumours without the network", in COSN 2014 - Proceedings of the 2014 ACM Conference on Online Social Networks, 2014, pp. 107-118.

[55] "Graph API", https://developers.facebook.com/docs/graph-api, accessed: 2018-07-26.

[56] Ugander, J., Karrer, B., Backstrom, L., Marlow, C., "The anatomy of the facebook social graph", arXiv:1111.4503, 2011.

[57] Wilson, C., Boe, B., Sala, A., Puttaswamy, K., Zhao, B., "User interactions in social networks and their implications", in Proceedings of the 4th ACM European Conference on Computer Systems, EuroSys'09, 2009, pp. 205-218.

[58] Viswanath, B., Mislove, A., Cha, M., Gummadi, K., "On the evolution of user interaction in facebook", in SIGCOMM 2009 - Proceedings of the 2009 SIGCOMM Conference and Co-Located Workshops, Proceedings of the 2nd ACM Workshop on Online Social Networks, WOSN 2009, 2009, pp. 37-42.

[59] Eckles, D., Kizilcec, R., Bakshy, E., "Estimating peer effects in networks with peer encouragement designs", Proceedings of the National Academy of Sciences of the United States of America, Vol. 113, No. 27, 2016, pp. 7316-7322.

[60] McAuley, J., Leskovec, J., "Learning to discover social circles in ego networks", in Advances in Neural Information Processing Systems, Vol. 1, 2012, pp. 539-547.

[61] Aral, S., Walker, D., "Identifying influential and susceptible members of social networks", Science, Vol. 337, No. 6092, 2012, pp. 337-341.

[62] Kosinski, M., Stillwell, D., Graepel, T., "Private traits and attributes are predictable from digital records of human behavior", Proceedings of the National Academy of Sciences of the United States of America, Vol. 110, No. 15, 2013, pp. 5802-5805.

[63] Bohn, A., Buchta, C., Hornik, K., Mair, P., "Making friends and communicating on facebook: Implications for the access to social capital", Social Networks, Vol. 37, No. 1, 2014, pp. 29-41.

[64] Jalali, M., Ashouri, A., Herrera-Restrepo, O., Zhang, H., "Information diffusion through social networks: The case of an online petition", Expert Systems with Applications, Vol. 44, 2016, pp. 187-197.

[65] "Facebook platform policy", https://developers.facebook.com/policy, accessed: 2018-07-26.

[66] Piškorec, M., Šmuc, T., Šikić, M., “Disentangling sources of influence in online social networks”, arXiv:1811.10372, 2018.

[67] Blondel, V., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., “Fast unfolding of communities in large networks”, Journal of Statistical Mechanics: Theory and Experiment, Vol. 2008, No. 10, 2008.

[68] Csardi, G., Nepusz, T., “The igraph software package for complex network research”, InterJournal, Vol. Complex Systems, 2006, pp. 1695.

[69] Goodman, L. A., “Snowball sampling”, Annals of Mathematical Statistics, Vol. 32, No. 1, 1961, pp. 148–170.

[70] Lee, N., Facebook nation: Total information awareness. Springer, 2014, Vol. 9781493917402.

[71] Schneble, C., Elger, B., Shaw, D., “The cambridge analytica affair and internet-mediated research”, EMBO Reports, Vol. 19, No. 8, 2018.

[72] “Cambridge analytica controversy must spur researchers to update data ethics”, Nature, Vol. 555, No. 7698, 2018, pp. 559-560.

[73] Tarran, B., “What can we learn from the facebook—cambridge analytica scandal?”, Significance, Vol. 15, No. 3, 2018, pp. 4-5.

[74] Isaak, J., Hanna, M., “User data privacy: Facebook, cambridge analytica, and privacy protection”, Computer, Vol. 51, No. 8, 2018, pp. 56-59.

[75] Verma, I., “Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks.”, Proceedings of the National Academy of Sciences of the United States of America, Vol. 111, No. 29, 2014, pp. 10779, cited By 4.

[76] Zimmer, M., “"but the data is already public": On the ethics of research in facebook”, Ethics and Information Technology, Vol. 12, No. 4, 2010, pp. 313-325.

[77] Narayanan, A., Shmatikov, V., “Myths and fallacies of "personally identifiable information"”, Commun. ACM, Vol. 53, No. 6, Jun. 2010, pp. 24–26.

[78] Vosoughi, S., Roy, D., Aral, S., “The spread of true and false news online”, Science, Vol. 359, No. 6380, 2018, pp. 1146-1151.

[79] Walker, D., Muchnik, L., “Design of randomized experiments in networks”, Proceedings of the IEEE, Vol. 102, No. 12, 2015, pp. 1940-1951.

[80] Jackman, M., Kanerva, L., “Evolving the irb: Building robust review for industry research”, Washington and Lee Law Review Online, Vol. 72, 2016.

[81] Daley, D., Kendal, D., “Stochastic rumors”, J. Inst. Maths Applics 1, p42., 1965.

[82] Maki, D., “Mathematical models and applications, with emphasis on social, life, and management sciences”, Prentice Hall., 1973.

[83] Hill, A., Rand, D., Nowak, M., Christakis, N., “Infectious disease modeling of social contagion in networks”, PLoS Computational Biology, Vol. 6, No. 11, 2010.

[84] Brauer, F., Castillo-Chavez, C., Mathematical Models in Population Biology and Epidemiology. Springer Verlag, 2012.

[85] Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J., Vespignani, A., “Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions”, PLoS Med, Vol. 4, No. 1, 01 2007.

[86] Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J. J., Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W., Colizza, V., Vespignani, A., “Seasonal transmission potential and activity peaks of the new influenza a(h1n1): a monte carlo likelihood analysis based on human mobility”, BMC medicine, Vol. 7, No. 45, 2009.

[87] Ugander, J., Backstrom, L., Marlow, C., Kleinberg, J., “Structural diversity in social contagion”, Proceedings of the National Academy of Sciences of the United States of America, Vol. 109, No. 16, 2012, pp. 5962-5966.

[88] Mahmoodi, A., Bahrami, B., Mehring, C., “Reciprocity of social influence”, Nature Communications, Vol. 9, No. 1, 2018.

[89] La Fond, T., Neville, J., “Randomization tests for distinguishing social influence and homophily effects”, in Proceedings of the 19th International Conference on World Wide Web, WWW ’10, 2010, pp. 601-610.

[90] Jacob Goldenberg, E. M., B. Libai, “Talk of the network: A complex systems look at the underlying process of word-of-mouth”, Marketing Letters, pp. 211–223, 2001.

[91] Narasimhan, H., Parkes, D. C., Singer, Y., “Learnability of influence in networks”, in Advances in Neural Information Processing Systems 28, Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., (ur.). Curran Associates, Inc., 2015, pp. 3186–3194, Available: http://papers.nips.cc/paper/5989-learnability-of-influence-in-networks.pdf

[92] Wang, C., Chen, W., Wang, Y., "Scalable influence maximization for independent cascade model in large-scale social networks", Data Mining and Knowledge Discovery, Vol. 25, No. 3, 2012, pp. 545-576, cited By 50.

[93] Porter, M. A., Gleeson, J. P., "Dynamical systems on networks: A tutorial", arXiv:1403.7663, 2014.

[94] Snijders, T., van de Bunt, G., Steglich, C., "Introduction to stochastic actor-based models for network dynamics", Social Networks, Vol. 32, No. 1, 2010, pp. 44-60.

[95] Dodds, P., Watts, D., "A generalized model of social and biological contagion", Journal of Theoretical Biology, Vol. 232, No. 4, 2005, pp. 587-604.

[96] Moreno, Y., Nekovee, M., Pacheco, A., "Dynamics of rumor spreading in complex networks", Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, Vol. 69, No. 6 2, 2004, pp. 066 130-1-066 130-7.

[97] McCullen, N. J., Rucklidge, A. M., Bale, C. S. E., Foxon, T. J., Gale, W. F., "Multiparameter models of innovation diffusion on complex networks", SIAM Journal on Applied Dynamical Systems, Vol. 12, No. 1, 2013, pp. 515-532.

[98] Melnik, S., Ward, J., Gleeson, J., Porter, M., "Multi-stage complex contagions", Chaos, Vol. 23, No. 1, 2013.

[99] Pérez-Reche, F., Ludlam, J., Taraskin, S., Gilligan, C., "Synergy in spreading processes: From exploitative to explorative foraging strategies", Physical Review Letters, Vol. 106, No. 21, 2011.

[100] Castellano, C., Fortunato, S., Loreto, V., "Statistical physics of social dynamics", Reviews of Modern Physics, Vol. 81, No. 2, 2009, pp. 591-646.

[101] Piškorec, M., Antulov-Fantulin, N., Miholić, I., Šmuc, T., Šikić, M., "Modeling peer and external influence in online social networks: Case of 2013 referendum incroatia", in Complex Networks & Their Applications VI, Cherifi, C., Cherifi, H., Karsai, M., Musolesi, M., (ur.). Cham: Springer International Publishing, 2018, pp. 1015–1027.

[102] Takaguchi, T., Masuda, N., Holme, P., "Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics", PLoS ONE, Vol. 8, No. 7, 2013.

[103] Karimi, F., Holme, P., "Threshold model of cascades in empirical temporal networks", Physica A: Statistical Mechanics and its Applications, Vol. 392, No. 16, 2013, pp. 3476-3483.

[104] Murphy, K. P., Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.

[105] Bezáková, I., Kalai, A., Santhanam, R., "Graph model selection using maximum likelihood", in ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning, Vol. 2006, 2006, pp. 105-112.

[106] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., Ghahramani, Z., "Kronecker graphs: An approach to modeling networks", Journal of Machine Learning Research, Vol. 11, 2010, pp. 985-1042.

[107] Schöning, U., "Graph isomorphism is in the low hierarchy", Journal of Computer and System Sciences, Vol. 37, No. 3, Dec. 1988, pp. 312–323.

[108] Papadopoulos, F., Kitsak, M., Serrano, M., Boguñá, M., Krioukov, D., "Popularity versus similarity in growing networks", Nature, Vol. 489, No. 7417, 2012, pp. 537-540.

[109] Barzel, B., Liu, Y.-Y., Barabási, A.-L., "Constructing minimal models for complex system dynamics", Nature Communications, Vol. 6, 2015.

[110] Stumpf, M., Porter, M., "Critical truths about power laws", Science, Vol. 335, No. 6069, 2012, pp. 665-666.

[111] Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A., "Microscopic evolution of social networks", in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 462–470.

[112] C.F., L., The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press, 2004.

[113] Medo, M., "Statistical validation of high-dimensional models of growing networks", Physical Review E, Vol. 89, 2014.

[114] Aliakbary, S., Motallebi, S., Rashidian, S., Habibi, J., Movaghar, A., "Noise-tolerant model selection and parameter estimation for complex networks", Physica A: Statistical Mechanics and its Applications, Vol. 427, 2015, pp. 100–112.

[115] Holland, P., Laskey, K., Leinhardt, S., "Stochastic blockmodels: First steps", Social Networks, Vol. 5, No. 2, 1983, pp. 109-137.

[116] Rosvall, M., Bergstrom, C., "An information-theoretic framework for resolving community structure in complex networks", Proceedings of the National Academy of Sciences of the United States of America, Vol. 104, No. 18, 2007, pp. 7327-7331.

[117] Peixoto, T., “Parsimonious module inference in large networks”, Physical Review Letters, Vol. 110, No. 14, 2013.

[118] Peixoto, T., “Hierarchical block structures and high-resolution model selection in large networks”, Physical Review X, Vol. 4, No. 1, 2014.

[119] Myers, S., Leskovec, J., “On the convexity of latent social network inference”, in Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010, 2010.

[120] Gomez-Rodriguez, M., Balduzzi, D., Schölkopf, B., “Uncovering the temporal dynamics of diffusion networks”, in Proceedings of the 28th International Conference on Machine Learning, ser. ICML ’11, 2011, pp. 561-568.

[121] Gomez-Rodriguez, M., Leskovec, J., Krause, A., “Inferring networks of diffusion and influence”, ACM Transactions on Knowledge Discovery from Data, Vol. 5, No. 4, 2012.

[122] Gomez Rodriguez, M., Leskovec, J., Schölkopf, B., “Structure and dynamics of information pathways in online media”, in WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining, 2013, pp. 23-32.

[123] Barzel, B., Barabási, A.-L., “Universality in network dynamics”, Nature Physics, Vol. 9, No. 10, 2013, pp. 673-681.

[124] Gleeson, J., “Binary-state dynamics on complex networks: Pair approximation and beyond”, Physical Review X, Vol. 3, No. 2, 2013.

[125] Singer, P., Helic, D., Taraghi, B., Strohmaier, M., “Detecting memory and structure in human navigation patterns using markov chain models of varying order”, PLoS ONE, Vol. 9, No. 7, 2014.

[126] Singer, P., Helic, D., Hotho, A., Strohmaier, M., “Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web”, in Proceedings of the 24th International Conference on World Wide Web, ser. WWW ’15, 2015, pp. 1003–1013.

[127] Strelioff, C. C., Crutchfield, J. P., Hübler, A. W., “Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling”, Physical Review E, Vol. 76, 2007.

[128] Guille, A., Hacid, H., "A predictive model for the temporal dynamics of information diffusion in online social networks", in WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion, 2012, pp. 1145-1152.

[129] Nash, S., "A survey of truncated-newton methods", Journal of Computational and Applied Mathematics, Vol. 124, No. 1-2, 2000, pp. 45-59.

[130] Holme, P., "Modern temporal network theory: a colloquium", European Physical Journal B, Vol. 88, No. 9, 2015.

[131] Crane, R., Sornette, D., "Robust dynamic classes revealed by measuring the response function of a social system", Proceedings of the National Academy of Sciences of the United States of America, Vol. 105, No. 41, 2008, pp. 15 649-15 653.

[132] Teng, X., Pei, S., Morone, F., Makse, H., "Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks", Scientific Reports, Vol. 6, 2016.

[133] Fawcett, T., "An introduction to roc analysis", Pattern Recognition Letters, Vol. 27, No. 8, 2006, pp. 861 - 874, rOC Analysis in Pattern Recognition.

[134] Grunwald, P., The Minimum Description Length Principle. The MIT Press, 2007.

[135] Bishop, C. M., Pattern Recognition and Machine Learning. Springer, 2007.

[136] Murray, I., Ghahramani, Z., "A note on the evidence and bayesian occam's razor", Gatsby Computational Neuroscience Unit, University College London, Tech. Rep., 2005, Available: http://mlg.eng.cam.ac.uk/zoubin/papers/05occam/occam.pdf

[137] Schwartz, G., "Estimating the dimension of a model", The Annals of Statistics, Vol. 6, No. 2, 1978, pp. 461-464.

[138] Akaike, H., "New look at the statistical model identification.", IEEE Transactions on Automatic Control, Vol. AC-19, No. 6, 1974, pp. 716-723.

[139] Rissanen, J., "Universal coding, information, prediction, and estimation.", IEEE Transactions on Information Theory, Vol. IT-30, No. 4, 1984, pp. 629-636.

[140] Rissanen, J., "Fisher information and stochastic complexity", IEEE Transactions on Information Theory, Vol. 42, No. 1, 1996, pp. 40-47.

[141] Vapnik, V. N., Statistical Learning Theory. Wiley-Interscience, 1989.

[142] Benjamini, Y., Hochberg, Y., "Controlling the false discovery rate: A practical and powerfull approach to multiple testing", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1, 1995, pp. 289-300.

[143] Stone, M., "Cross-Validatory Choice and Assessment of Statistical Predictions", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 36, No. 2, 1974, pp. 111–147.

[144] Page, L., Brin, S., Motwani, R., Winograd, T., "The pagerank citation ranking: Bringing order to the web", Available: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf 1999.

[145] Bonacich, P., "Power and centrality: A family of measures", American Journal of Sociology, Vol. 92, No. 5, 1987, pp. 1170-1182.

[146] Langville, A., Meyer, C., "A survey of eigenvector methods for web information retrieval", SIAM Review, Vol. 47, No. 1, 2005, pp. 135-161.

[147] Kleinberg, J., "Authoritative sources in a hyperlinked environment", Journal of the ACM, Vol. 46, No. 5, 1999, pp. 604-632.

[148] Srivastava, A., Chelmis, C., Prasanna, V., "The unified model of social influence and its application in influence maximization", Social Network Analysis and Mining, Vol. 5, No. 1, 2015, pp. 1-15.

[149] Popa, A., Frincu, M., Chelmis, C., "A distributed algorithm for the efficient computation of the unified model of social influence on massive datasets", in Proceedings - IEEE High Performance Extreme Computing Conference (HPEC). IEEE, 09 2017, pp. 1-7.

[150] Bovet, A., Makse, H., "Influence of fake news in twitter during the 2016 us presidential election", Nature Communications, Vol. 10, No. 1, 2019.

[151] Dhand, A., White, C., Johnson, C., Xia, Z., De Jager, P., "A scalable online tool for quantitative social network assessment reveals potentially modifiable social environmental risks", Nature Communications, Vol. 9, No. 1, 2018.

[152] Rutherford, A., Cebrian, M., Dsouza, S., Moro, E., Pentland, A., Rahwan, I., "Limits of social mobilization", Proceedings of the National Academy of Sciences, 2013.

[153] Yoshikawa, Y., Saito, K., Motoda, H., Ohara, K., Kimura, M., "Acquiring expected influence curve from single diffusion sequence", Lecture Notes in Computer Science

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 6232 LNAI, 2010, pp. 273-287.

[154] Du, N., Liang, Y., Balcan, M.-F., Song, L., "Influence function learning in information diffusion networks", in 31st International Conference on Machine Learning, ICML 2014, Vol. 5, 2014, pp. 4118-4135.

List of Figures

List of Tables

List of Algorithms

Abbreviations

AIC Akaike Information Criteria
API Application Programming Interface
AsIC Asynchronous Independent Cascade
AsLT Asynchronous Linear Threshold
AUC Area Under Curve
CV Cross-Validation
EM Expectation Maximization
EXP Exponential Decay
FAQ Frequently Asked Questions
FDR False Discovery Rate
HMM Hidden Markov Models
IC Independent Cascade
iid independent identically distributed
ISS Ignorant Spreader Stifler
LOG Logistic Threshold
LT Linear Threshold
MAP Maximum a Posteriori
MCMC Markov Chain Monte Carlo
MDL Minimum Description Length
ROC Receiver Operating Characteristic
SI Susceptible Infected
SIR Susceptible Infected Recovered
SIS Susceptible Infected Susceptible
SRM Structural Risk Minimization
T-BaSIC Time-Based Asynchronous Independent Cascades
VC Vapnik-Chervonenkis

Biography

Matija Piškorec was born in Bjelovar, Croatia, in 1986. He graduated from the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, in 2010, with a Masters in computing. In 2010 he was awarded with the Rector's Award of the University of Zagreb. From 2010 to 2011 he worked as a research intern at Max Perutz Laboratories at Vienna Biocenter, Vienna, Austria, in the group for Computational Biophysics of Bojan Žagrović. From 2011 to 2013 he was employed at Division of Electronics, Ruđer Boškovic Institute (RBI), Zagreb, as a project associate on EU FP7 projects "An e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science" (e-LICO) and "Forecasting Financial Crisis" (FOC II). From 2013 to 2019 he was employed as a research assistant on RBI. He enrolled at doctoral studies on FER in 2014, under joint supervision of Mile Šikić (FER) and Tomislav Šmuc (RBI). Currently he is employed as a young researcher at Scientific Center of Excellence, project "Datacross". During his career he performed several international research visits, including to ETH Zurich in the group for Computational Social Science of Dirk Helbing, and Aalto University, Helsinki, in the Data Mining Group of Aristides Gionis. Since 2013 he is a teaching assistant on a course "Machine Learning" on the Faculty of Science, Department of Mathematics, University of Zagreb. He also participated on several summer schools for graduate students in Croatia and abroad.

He has published 3 original peer-reviewed research papers in Q1 journals, and 6 research papers in international peer-reviewed conferences and conference proceedings.

List of Publications

Journal papers

1. † Piškorec, M., Šmuc, T., Šikić, M., *"Disentangling Sources of Influence in Online Social Networks"*, *IEEE Access*, Volume 7, Issue 1, pp. 131692-131704, Published September 12 2019, DOI: 10.1109/ACCESS.2019.2940762
2. Brbić M., Piškorec M., Vidulin V., Kriško A., Šmuc T., Supek F., *"The landscape of microbial phenotypic traits and associated genes"*, *Nucleic Acids Research*, 44(21):10074-10090, Published December 1 2016, DOI: 10.1093/nar/gkw964
3. Piškorec M., Antulov-Fantulin N., Kralj-Novak P., Mozetič I., Grčar M., Vodenska I. and Šmuc T., *"Cohesiveness in Financial News and its Relation to Market Volatility"*, Scientific Reports 4, 5038, Published May 22 2014, DOI: 10.1038/srep05038

Conference proceedings

1. Antulov-Fantulin N., Tolić D., Piškorec M., Ce Z., Vodenska I., *"Inferring Short-Term Volatility Indicators from the Bitcoin Blockchain"*, Complex Networks & Their Applications VII. COMPLEX NETWORKS 2018. Studies in Computational Intelligence, vol 813. Springer, Cham. DOI: 10.1007/978-3-030-05414-4_41
2. † Piškorec M., Antulov-Fantulin N., Miholić I., Šmuc T., Šikić M., *"Modeling Peer and External Influence in Online Social Networks: Case of 2013 Referendum in Croatia"*, Complex Networks & Their Applications VI. COMPLEX NETWORKS 2017. Studies in Computational Intelligence, vol 689. Springer, Cham. DOI: 10.1007/978-3-319-72150-7_82
3. Brbić M., Piškorec M., Vidulin V., Kriško A., Šmuc T., Supek F., *"Phenotype Inference from Text and Genomic Data"*, European Conference on Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science, vol 10536, pp. 373-377, Springer, Cham. DOI: 10.1007/978-3-319-71273-4_34
4. Piškorec M., Sluban B., Šmuc T., *"MultiNets: Web-Based Multilayer Network Visualization"*, European Conference on Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science, vol 9286. Springer, Cham. DOI: 10.1007/978-3-319-23461-8_34
5. Piškorec M., Bošnjak M. and Šmuc T., *"Meta-modeling Execution Times of RapidMiner Operators"*, Proceedings of the 3rd RapidMiner Community Meeting and Conference (RCOMM 2012)

† Related to the doctoral research.

6. Piškorec M., Antulov-Fantulin N., Ćurić J., Dragoljević O., Ivanac V. and Karlović L., *"Computer vision system for the chess game reconstruction"*, Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2011), pp. 870-876
7. Sović I., Antulov-Fantulin N., Čanadi I., Piškorec M. and Šikić M., *"Parallel Protein Docking Tool"*, Proceedings of the 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2010), pp. 1333 - 1338. **Best student paper award.**

Other publications and presentations

1. † Piškorec M., Antulov-Fantulin N., Miholić I., Šmuc T., Šikić M., *"Inference of influence in social networks"*, talk given at the 8^{th} Conference on Complex Networks (CompleNet 2017), 21^{st} - 24^{th} March 2017, Dubrovnik, Croatia
2. † Piškorec M., Antulov-Fantulin N., Miholić I., Šmuc T., Šikić M., *"Modeling peer and external influence in online social network"*, poster at the Network Science conference (NetSci 2015), 1^{st} - 5^{th} June 2015, Zaragoza, Spain
3. Piškorec M., Antulov-Fantulin N., Kralj-Novak P., Mozetič I., Grčar M., Vodenska I. and Šmuc T. *"Cohesiveness in Financial News and its Relation to Market Volatility"*, a talk given at the European Conference on Complex Systems (ECCS 2014), 22^{nd} - 26^{th} September 2014, Lucca, Italy
4. † Piškorec M., Antulov-Fantulin N., Miholić I., Šmuc T., Šikić M., *"Modeling voting activity dynamics in social network during December 1st 2013 referendum in Croatia"*, ignite talk presented at the European Conference on Complex Systems (ECCS 2014), 22^{nd} - 26^{th} September 2014, Lucca, Italy
5. Piškorec M., Antulov-Fantulin N., Šmuc T., Mozetič I., Grčar M., Kralj-Novak P. and Vodenska I., *"Quantifying the Impact of Cohesiveness in Financial News"*, a talk given at the European Conference on Complex Systems (ECCS 2013), Barcelona, Spain

Životopis

Matija Piškorec, mag. ing. comp., rođen je u Bjelovaru 1986. Diplomirao je računarstvo na Fakultetu elektrotehnike i računarstva (FER) Sveučilišta u Zagrebu 2010. Iste godine je dobio Rektorovu nagradu Sveučilišta u Zagrebu. Od 2010. do 2011. radi kao istraživački pripravnik na Max F. Perutz Laboratories u Beču u grupi za računalnu biofiziku pod vodstvom Bojana Žagrovića. Od 2011. do 2013. zaposlen je na Zavodu za elektroniku na Institutu Ruđer Bošković (IRB) kao projektni suradnik na nekoliko međunarodnih EU FP7 projekata - "An e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science" (e-LICO) and "Forecasting Financial Crisis" (FOC II). Od 2013. do 2019. zaposlen je kao asistent na IRB-u. Doktorski studij na FER-u je upisao 2014. pod dvojnim mentorstvom Mile Šikića s FER-a i Tomislava Šmuca s IRB-a. Trenutno je zaposlen kao mlađi istraživač na Znanstvenom centru izvrsnosti iz znanosti o podacima na projektu "Datacross". Njegovi istraživački interesi su u području strojnog učenja, znanosti o podacima i statističkog zaključivanja u kompleksnim sustavima. Tijekom svoje karijere boravio je na nekoliko međunarodnih istraživačkih institucija, uključujući ETH Zurich u grupi za Computational Social Science pod vodstvom Dirk Helbinga i Aalto University u Helsinkiju u grupi Data Mining Group pod vodstvom Aristidesa Gionisa. Od 2013. sudjeluje kao asistent na kolegiju "Strojno učenje" na Matematičkom odsjeku Prirodoslovno matematičkog fakulteta Sveučilišta u Zagrebu. Objavio je tri originalna istraživačka rada u časopisima Q1 kategorije i šest radova u zbornicima međunarodnih konferencija.