iSocial

Issue 3, May 2016

ISOCIAL: DECENTRALIZED ONLINE SOCIAL NETWORKS

28 - 29 January 2016: "iSocial Research Meeting" Milan, Italy

Inside this Issue

TRENDING TOPICS ON DECENTRALIZED ONLINE SOCIAL NETWORKS

A fully Decentralized P2P Publish/Subscribe Notification System

Publish/subscribe (pub/sub) mechanisms constitute an attractive communication paradigm in the design of large- scale notification systems for Online Social Networks (OSNs). In our research, we focus on designing a fully decentralised pub/sub notification system that leverages the social network dynamics in order to enhance the dissemination of the OSN notifications".

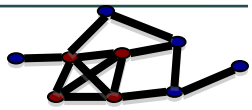
See Page 4

Privately Locating Nearby Users in Online Social Networks

Millions of online social network users often need to locate friends and contacts that are in their nearby vicinity. We propose a method for offering this functionality . Maintaining, at the same time, the privacy of users without reveling their exact location.

See Page 6

Finding Densest Subgraph for Sliding Window Model



The densest subgraph problem is a fundamental graph theory problem, which has various applications, like social network analysis, community detection, event detection, link spam detection, computational biology, distance query indexing, etc. This problem becomes challenging when dealing with dynamic streams, where the graph is changing and users are interested in only the recent data points. This calls for an efficient algorithm for densest subgraph in the sliding window model that enables extracting densest subgraph in real time for the recent input stream.

See Page 7

Gossip-based Behavioral Group Identification in Decentralized OSNs

Discovery of meaningful groups of users that share the same behavioral patterns in DOSN, where there is no central infrastructure, is challenging. In the fully distributed social graph, each user can only communicates with his/her direct friends without sending all the private information in a raw form. We propose a fully decentralized clustering algorithm (Newscast EM), a probabilistic gossip-based randomized communication approach, originally developed for clustering users in peer-to-peer networks and then, making Newscast EM to be applicable on top of the DOSNs and apply it to identify behavioral group of users.

See Page 8

Multi-party access control for online social networks

The existing mechanisms employed in online social networks do not allow all the users associated with a resource to specify how this resource should be distributed in the network. In general, the users related to a resource are exposed to the access control decisions of the uploader, which may not be privacy concerned. In our research we design a multi-party access control model that allows all the associated users to contribute on the specification and enforcement of the

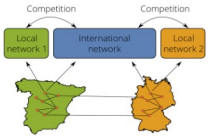
See Page 10

What is the destiny of social networking sites data after the data owners pass away?

The management of social networks post-mortem data is a critical topic that could be subject for investigation within multiples disciplines, such as legacy management by law, human rights, privacy preservation, technological tools, etc. From a high level view, there is the challenge of balancing between preserving users privacy and security preferences even after they have gone, and answering the needs of the deceased user's families and friends. Moreover, there is the issue of whether all the generated content is to be simply thrown to nonexistence, or it might hold precious contributions to the intellectual or social arenas, that would need to be commemorated, saved, or exploited for scientific, intellectual, or social benefits. Our research investigates such issues under a decentralized OSN model, where their management is technically more challenging.

See Page 12

Competition between global and local online social networks



"We study the impact of heterogeneity in network fitnesses on the competition between an international network, such as Facebook, and local services. The higher fitness of international networks is induced by their ability to attract users from all over the world, which can then establish social interactions without the limitations of local networks. Our findings shed new light on the overtake of Facebook at the cost of many local networks."

See Page 14

Users Key Locations in Online Social Networks: Identification and Applications

Cities around the world have the need to frequently monitor the commuting patterns of their inhabitants in order to maintain and improve the quality of their transportation systems. With this project we aim in identifying users key locations and cities overall transportation patterns using data retrieved from Online Social Networks. Furthermore, we investigate the influence of key locations in user's Online Social Networking activity.

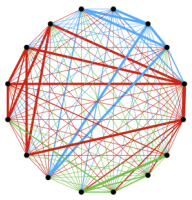
See Page 16

Building a Spam Free Social Network

As Online Social Networks continue to grow in popularity, a spam marketplace has emerged that includes services selling fraudulent accounts, as well as a nucleus of spammers who propagate large-scale spam campaigns. Differently from existing works, our framework employs unsupervised machine learning algorithm, hence diminishes both of the training cost and the need of labelled training data. Furthermore, our proposed framework can timely detect the spam in large-scale social networks by analyzing and clustering users' behaviors in order to detect the spammers.

See Page 17

Large Scale Topic Detection using Node-Cut Partitioning on Dense



Topic Detection in Text (TDT) is the problem of automatic determination of mutually related documents in a given corpus of text. Extracting topics in online social networks (OSNs) is a challenging problem due to fast speed of publication and large number of noisy messages that respectively affect scalability and quality of proposed solutions. We proposed a solution for TDT to overcome those limitations by combining Dimensionality Reduction (DR) with graph analytics. We use a DR method called Random Indexing (RI) and a partitioning algorithm inspired from a method called JaBeJa-VC developed in our group

See Page 20

Individualism and Collectivism in Social Dynamics

The structure and dynamics of Online Social Networks (OSNs) have been thoroughly studied. Existing methods allow to measure metrics that indicate how sparse or dense is a network what is the role of central hubs and how these properties evolve over time.

See Page 23

A Fully Decentralized P2P Publish/Subscribe Notification System

One of the fundamental services of social networks is the real-time delivery of notifications to the social users. Notifications constitute one of the primary way social users first learn about content that their social friends publish or their preferable sources (e.g. groups, pages) share. Publish/Subscribe (Pub/Sub) systems are an attractive solution for the design of large-scale social notification systems. However, such Pub/Sub systems require thousands of servers (brokers) placed at strategic points across the globe in order to achieve the desired scalability. For example, IBM utilizes over a thousand of servers on geographically distributed data centers in order to deliver the tennis match scores to millions of users around the world. Moreover, Akamai maintains an infrastructure with more than 175.000 dedicated servers across 100 countries in order to deliver hundreds of billions of Internet interactions daily. Taking into account the growth of the IoT and its integration with the social networks, the number of resources required to deliver a Pub/Sub system increases and motivates for a more advanced designs.

Fully decentralised publ sub notification system that leverages the social network dynamics in order to enhance the dissemination of the OSN notifications

The above motivation attracted the attention of both the research community and industry to design and deploy topic-based pub/sub systems over peer-to-peer (P2P) environments. A key characteristic of the P2P Pub/Sub implementations is that they leverage the P2P network and concentrate on the design of the Pub/Sub routing process without any further development on the P2P infrastructure. While this is easy to implement, such P2P Pub/Sub systems present high traffic overhead, since peers have to receive and forward messages that are not interested; these peers are also known as relay nodes.

In our research work, we emphasize on the construction of the P2P network, rather than focusing on the P2P Pub/Sub routing process, in order to enhance the notification dissemination. We form our P2P Pub/Sub system using a three-layer architecture. In our designed P2P Pub/Sub system, we organize the peers in the P2P network by leveraging the social network friendships. Specifically, two users that are friends in the social network are more likely to be also connected in the P2P network, and thus the notification dissemination will be accomplished without any relay node. Moreover, our results suggest that using this approach we manage to reduce the total number of messages required to propagate a message to all the subscribers.

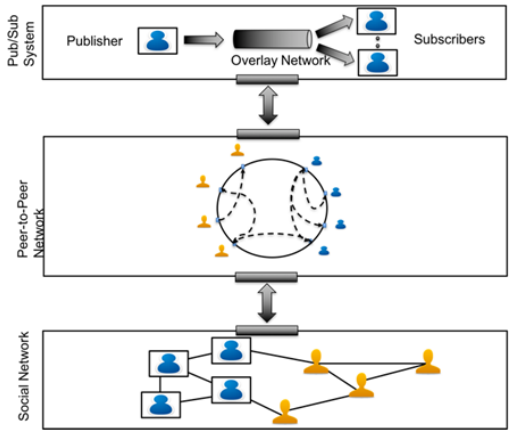
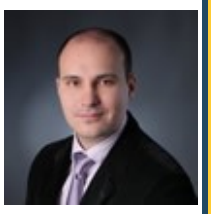


Figure 1 - Three-layer architecture of our designed P2P Pub/Sub System

Indeed, modern social networks, such as Facebook and Twitter, can utilize our P2P Pub/Sub system to offload processing from their dedicated resources and provide a real-time delivery notification system. In addition, services which already exploit the P2P network to provide a Pub/Sub system, such as Spotify, can benefit from our research by reducing the high traffic overhead that the relay nodes pose.

Stefanos Antaris
antaris.stefanos@cs.ucy.ac.cy
ESR iSocial Fellow
University of Cyprus (UCY), Cyprus



Privately Locating Nearby Users in Online Social Networks

Online social networks (OSNs) are among the most popular services of the World Wide Web. Their millions of users often need to locate friends and contacts that are in their nearby vicinity. This, however, requires revealing their exact locations, compromising their privacy. To this end, we propose a computationally efficient method for locating nearby users of online social networks using publically available information, such as public points of interest and public social network events, as points of reference.

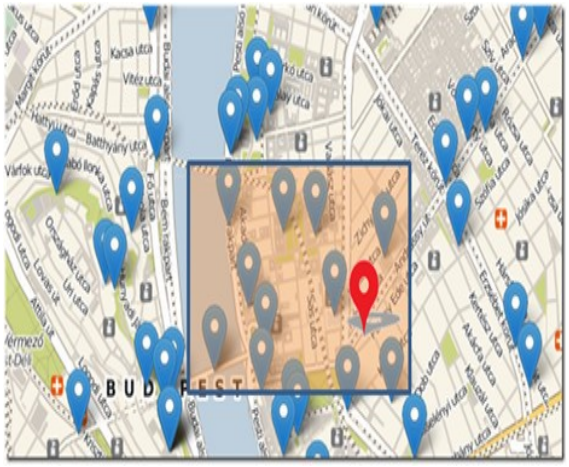
In this project, we propose associating users with publicly available information for discovering their neighbors without compromising their privacy

Users in online social networks (OSNs) and users of Location Based Services (LBSs) often need to interact with friends or peers in their vicinity. However, this either requires revealing their exact location to the respective service provider, or using computationally intensive solutions based on cryptographic algorithms. Such approaches are not suitable in the context of OSNs, in cases where users' privacy preservation is a primary concern

Privately locating nearest neighbors of users, is a known problem of Location Based Services, referred to in the respective literature as the privacy preserving k – nearest neighbors problem. Most solutions adhere to computationally expensive cryptographic methods, approaches that lead to reduced scalability and require significant computational resources, considering the continuously growing number of users of online social networks .

In this project, we propose associating users with publicly available information, as an intermediate point of comparison, for discovering their neighbors without compromising their privacy. This information may be publicly available GPS Point-Of-interest (POI) lists, OSN geotagged events, etc. User proximity is going to be assessed based on the common points of interest specific users are associated with.

To avoid triangulation, we are going to use a bounding box or circle around the user's location and move each time its center by a random numerical factor resulting from a random distribution, so that, each time a user wishes to find her nearby friends, a new cycle, or bounding box is calculated and a correlated POI is selected within the specific radius or bounding box.



We aspire that we will able to offer approximately correct results, with significantly less computational resources, a fact that is greatly applicable to the online social networks use case, with millions of users. Additionally, we expect be able to gather collective information regarding users' trends and behaviors, without, however, compromising their privacy.

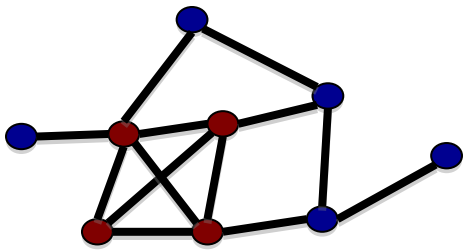
Alexandros Karakasidis
akarakasidis@cs.ucy.ac.cy
ER iSocial Fellow
University of Cyprus (UCY), Cyprus

Finding Densest Subgraph for Sliding Window Model

The densest subgraph problem is a fundamental graph theory problem, which has various applications, like social network analysis, community detection, event detection, link spam detection, computational biology, distance query indexing, etc. Any algorithm to solve densest subgraph problem aims to find a subgraph with the maximal density, where density is defined as a ratio between number of edges and nodes. There are also many other definitions of communities including cliques, quasi-cliques, α-β-communities, and k-cores. Also, similar definitions are often extended to find set of *top-k* subgraphs, where subgraphs might be node-disjoint, edge-joint or overlapping.

Densest Subgraph in Sliding Window Model

These subgraphs enable extracting useful information from a graph. In social network context, densest subgraphs might correspond to communities, i.e., sets of users sharing similar interests or being affiliated with a same organization such as a university or a company. Similarly, entities such as city, person and company names, starting suddenly to co-occur in tweets might indicate the occurrence of some interesting event involving the corresponding entities. Telecom providers use dense subgraph algorithm to figure out the region of high users interactions. Finding dense subgraphs might help in identifying interesting and novel patterns in gene annotation networks. In the past researchers have used the densest subgraph for link spam detection.



In the aforementioned application scenarios, data is large and inherently dynamic. For example, in Facebook, users join and leave the social network frequently, with new friendship links being established or removed all the time. In Twitter, tweets are generated incessantly making older tweets less interesting. As a result, communities evolve overtime; new events trigger new dense subgraphs in the corresponding entity relationship graph, while changing distances between nodes in a graph requires frequent re-indexing. For this reason, many data analyst are often interested in only the recent and fresh data (e.g., every month, every day, every hour, etc). This calls for an efficient algorithm for densest subgraph in the sliding window model that enables extracting densest subgraph in real time for the recent input stream.

Muhammad Anis Uddin Nasir
anisu@kth.se
ESR iSocial Fellow
Royal Institute of Technology (KTH), Sweden

Gossip-based Behavioral Group Identification in Decentralized OSNs

Decentralized online social networks (DOSNs) are distributed systems providing social networking services and become extremely popular in recent years. In DOSNs, the aim is to give the users control over their data and keeping data locally to enhance privacy. Discovery of meaningful groups of users that share the same behavioral patterns in social networks has became an active research area. Behavioral group identification has many valuable applications. For instance, it can be helpful for improving recommendation systems, it can be used for advertisement purposes, direct marketing, and for risk assessment in online social networks. The key idea in risk assessment is that the more the target user's behavior diverges from those of other similar users, the more the target user is risky. Therefore, risk assessment approaches require to identify similar users that share the same behavioral patterns.

The goal is identifying behavioral group of users in the fully distributed social graph, where each user can only communicates with his/her direct friends without sending all the private information in a raw form

By considering a social network as a graph, the main problem is that all users are connected in friend to friend graph, but users that share the same behavioral patterns not necessarily have friendships in the graph. Furthermore, in investigating the discovery of behavioral groups, we have cast our attention to DOSNs. Therefore, there is no central infrastructure and the discovery of behavioral groups is more challenging than in the centralized setup. In the fully distributed social graph, each user can only communicates with his/her direct friends without sending all the private information to his/her direct friends in a raw form.

The problem of finding similar users in social networks has been widely studied in the context of community detection. Those community detection approaches that are pure link-based, relying on topological structures, fail to group users with the same behavioral patterns in that such users might belong to different communities based on their friendship links. Moreover, some of the community detection approaches are content-based that is relying on the analysis of the content generated by each user. But, the major problem of these approaches is the overhead of graph constraction based on similarity measures, that is not suitable for real-time applications. However, there are some stream-based community detection methods suitable for real-time applications. But, most of these approaches are link-based, and they do not consider the personal feature vector of users.

To alleviate the limitations of existing approaches, we propose a fully decentralized clustering algorithm which is capable of clustering distributed information without requiring central control. The selected clustering model requires specific aspects to be considered such as: the final clustering model should maintain a reasonable performance compared to a centralized clustering model and should be robust in that it does not fail easily when some of the users leave the network or do not answer messages. Also, all users should be able to have the final clustering model at any time after convergence in order to assign a group to themselves and also to their direct friends in a local way.

Finally, we need to minimize the communication cost by decreasing the number of messages and the size of them as well. These requirements bring us to exploit Newscast EM, a probabilistic gossip-based randomized communication clustering approach, originally developed for clustering users in peer-to-peer networks. In Newscast EM, each user initializes a local estimation of the parameters of the clustering model and then, contacts a random user from all users in the network, to exchange his/her parameters estimation and aggregate them by weighted averaging. The choice of random selection is crucial to the wide dissemination of the gossip, since, the probability of a user being sampled is proportional to his/her degree.

Gossip based peer-sampling service provides a user with a uniform random subset of all users in the peer-to-peer network. But, the main difference between peer-to-peer and social networks is that in peer-to-peer networks each user can directly communicate with any other user in the network to exchange information. On the contrary, in social networks, each user can just communicate directly with his/her direct friends. Therefore, we use the random-sampling implementation for social network. The main contribution of this work is making Newscast EM to be applicable on top of DOSNs and apply it to identify behavioral group of users. Our goal is to achieve a comparable accuracy compared to a centralized scheme. The advantages of this distributed behavioral group identification are: 1. the usage of both social and individual patterns of users, 2. feature values of users are never send over the network in a raw form and 3. it has low computation and communication cost. In order to evaluate our approach, we implement our algorithm and test it in a real Facebook graph.

Laleh Naeimeh
Naimeh.laleh@gmail.com
ESR iSocial Fellow
University of Insubria (INSUB), Italy

Multi-party Access Control for Online Social Networks

"Users can preserve their privacy by controlling the way resources are shared in the network"

The volume and nature of the data available in online social networks raises alarm regarding users' privacy. Generally, users can preserve their privacy by controlling the way resources are shared in the network. But, despite the efforts by service providers and the research community to design effective access control mechanisms, users cannot fully control resources published by others.

Typically, the mechanisms employed in online social networks consider the uploader of a resource as owner, but not the users related to that resource as co-owners. This results on a lack of control from those users that are associated, in some way, with that resource. Therefore, those users are exposed to the access control decision of the uploader, which may not be privacy concerned.

Several approaches and mechanisms have been proposed in the literature for solving the problem of privacy conflicts in collaborative multi-user environments. Unfortunately, all these approaches either rely on a *trusted* service provider to solve cases of privacy conflicts and to enforce a mutually accepted access control policy, or they assume that the uploader and associated users play in an honest way, without the intention to enforce their own privacy preferences over those of the other associated users.

In our research we design a multi-party access control model that allows all the users associated with a resource to contribute on the specification and enforcement of the access control policy

In our work we do not consider the social network as *fully trusted* for accessing the data, resolving privacy conflicts between the users, and enforcing the access control policy. The online social network is only trusted to provide the correct (encrypted) resources, when requested. Furthermore, we consider that the data uploader and the associated users have the intention, and possibly try, to enforce their own privacy preferences over those of the other users, but always follow the protocol *honestly*, as malicious behavior can be easily detected.

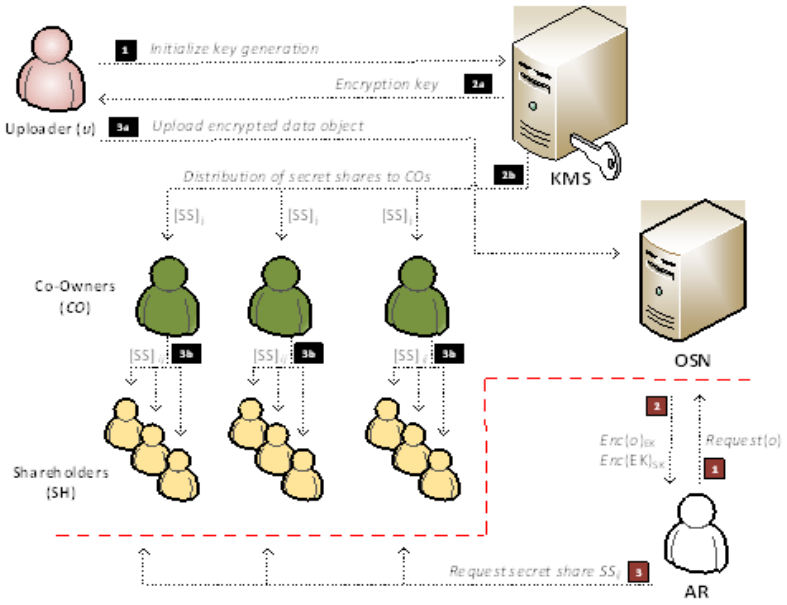


Figure 1 - {Title}

In order to solve the problem of privacy conflicts, we design a collaborative multi-party access control model that allows all the users associated with a resource to participate on the specification of the access control policy. This model determines that the data is encrypted before being uploaded to the online social network, and that a threshold-based secret sharing scheme is used for creating a number of tokens (secret shares) from the encryption key. After that, the tokens are distributed to the associated users' trusted friends (shareholders), which are set responsible for enforcing the access control policy, by providing the share they hold to the access requesting users.

According to this model, a user requesting access to the data has to convince a sufficient number of shareholders to provide the share they hold (according to the threshold), for being able to re-generate the encryption key, for decrypting the data.

Panagiotis Ilia
pilia@ics.forth.gr
ESR iSocial Fellow,
Foundation for Research and Technology-Hellas (FORTH) , Greece

What is the Destiny of Social Networking Sites Data after the Data Owners Pass Away?

"By the end of this year, there'll be nearly a billion people on this planet that actively use social networking sites. The one thing that all of them have in common is that they're going to die. While that might be a somewhat morbid thought, I think it has some really profound implications that are worth exploring."

This is how Adam Ostrow, an online journalist and Editor in Chief at Mashable, opened his five minute TED Talk titled "After your final status update."[1] The talk was filmed about five years ago, in 2011, leaving us with the question: how many of those billion active users of social networking sites have died since then? In fact, a more interesting question is: what has been the destiny of the left behind data of those who have died?

Very recently this year, the Independent online news-site published an article claiming that Facebook will have more dead users than alive ones by the end of the century.[2] The article based on reported results of a research study that drew the conclusion based on modeling the number of users joining Facebook against those who are dying.

In our research, we approach this problem of post-mortem data management from a data value and a data function perspectives, whilst ensuring the privacy preservation of users expressed preferences

The management of social networks post-mortem data is a critical topic that could be subject for investigation within multiples disciplines, such as legacy management by law, human rights, privacy preservation, technological tools, etc. From a high level view, there is the challenge of balancing between preserving users privacy and security preferences even after they have gone, and answering the needs of the deceased user's families and friends. Moreover, there is the issue of whether all the generated content is to be simply thrown to nonexistence, or it might hold precious contributions to the intellectual or social arenas, that would need to be commemorated, saved, or exploited for scientific, intellectual, or social benefits.

Out of the major existing social networking sites, Facebook has been the one to give considerable interest to the management of its dead users accounts. After a series of changes, driven by cases of reported incidents and requests from the families of deceased users, Facebook settled on the idea of memorializing the accounts of its dead users. A memorial account is a frozen one, to which login is disabled, and from which no further action or content could be generated. Starting from about a year ago, users, whilst still alive, have the possibility to identify a *legacy contact* to their accounts. The legacy contact, if specified in the account's settings, will have the possibility to make a final status update on behalf of the deceased user, could make a last change to the profile's picture, and could have the possibility to download an archive of the account's pictures and posts. Depending on the account's privacy settings, friends could continue to share posts (memories) on the memorialized account's timeline.

All the data in the account remains accessible as per the audiences it was initially posted for, except from private messages and conversations that would no more be accessible. Very recently, the option of choosing to have the account completely deleted upon passing away is available as a privacy setting that users could customize during their life. Another online giant that provides a diverse portfolio of online socializing services, Google, could be considered one of the first to design and offer a solution for post mortem data management of its users accounts.

1. https://www.ted.com/talks/adam_ostrow_after_your_final_status_update?language=en#t-6834
2. http://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-will-have-more-dead-people-than-living-ones-by-the-end-of-the-century-researcher-claims-a6917411.html

In 2013, Google introduced a feature called the *Inactive Account Manager*. Google users can select up to 10 trusted contacts from their contacts list to entrust them with their data, or selected portions of it, should they become unable to manage and use their accounts (mostly because of long inactivity that would refer to death).

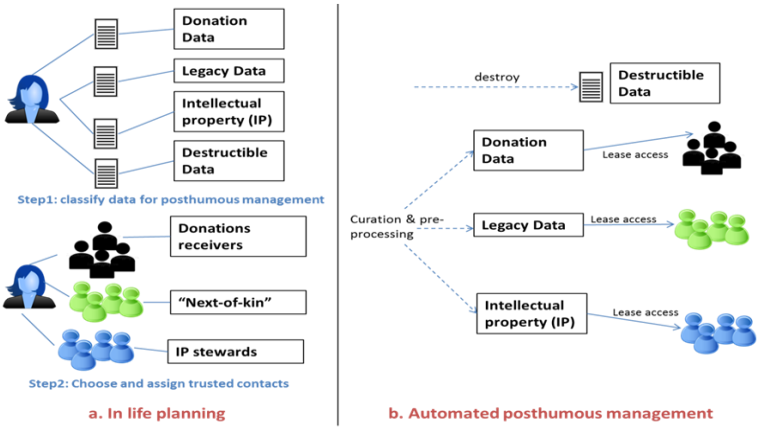
Users can set the length of their inactivity period beyond which Google will execute their pre-prepared plan regarding their chosen trusted contacts and the specified data that they would like to transfer to them.

These features, being better than having the account with all its related data simply expire, still suffer from some limitations. For instance, users can choose to have their data, or selected portions of it, sent to their nominated trusted contacts; however, they cannot put any restriction on what could be done with this data or on how it could be utilized. For instance, if a Google user has generated number of videos and published them on Youtube, once access and control over these videos is passed to their trusted contacts they become their owners. This might violate intellectual property rights. The same could apply for Facebook pages that users could create to run a personal business, to advertise for a talent (such as poetry, painting, etc.), etc. Herein, comes the debate on whether social network's data should be treated as inheritance data or as content that requires stewardship style of management without full transfer of ownership.

In our research, we approach this problem of post-mortem data management from a data value and a data function perspectives, whilst ensuring the privacy preservation of users expressed preferences. We claim that the content users share on social networking sites should not be treated with the same mechanisms, but rather controlled based on its value and function. Specifically, we differentiate between four categories of post mortem data functions that content generated in a social networking site could be considered under: 1. donation data, 2. legacy data, 3. intellectual property, and 4. destructible data.

Our research aims to design technical tools to annotate users data, based on their expressed preferences, and classify it under one of the suggested functional categories.

Our objective is to design an integrated framework within which users could be assisted to create their *posthumous digital data plans*, and that would ensure the execution, data dissemination, and follow-up on behalf of users after they pass away.



Another future challenge in our research plan is regarding the provision of a similar framework under a decentralized architectures for social networking sites. In the absence of a central data storage and management authority, the problem opens richer research challenges.

Figure 1 - Our functional approach for post-mortem data planning and management

Leila Bahri
Leila.Bahri@uninsubria.it
ESR iSocial Fellow
University of Insubria (INSUB), Italy

Competition Between Global and Local Online Social Networks

Understanding the complex dynamics of the digital world constitutes an important challenge for interdisciplinary science. Online social networks (OSNs) constantly compete to attract and retain users' attention. To meet this challenge, we describe the web as a complex, digital ecosystem in which interacting networks play the role of species in competition for survival.

First, we introduce a very general and concise theory of the digital ecosystem. Akin to standard ecological theories of competing species, the attractiveness of OSNs increases with their performance following a preferential attachment (or rich-get-richer) mechanism. However, unlike the case of standard ecology, the total amount of users' attention is a conserved quantity, which induces diminishing returns in the attractiveness of each network. Over a range of parameters, the combination of these two mechanisms leads to stable states of coexistence of several networks, in stark contrast to the competitive exclusion principle.

Is bigger always better? Not necessarily. In the digital world, local online social networks could have dominated and the overtake of Facebook could not have taken place, our model suggests

In particular, we investigate the competition between local networks operating in single countries and one international network that operates in all countries. Therefore, a proper description of this system must necessarily involve the network of worldwide social interactions between different countries (see Fig. 1). We show that the effect of inter-country social ties can be mapped to an increased fitness of the international network by means of an effective activity. Interestingly, there is a critical global coupling strength below which networks can coexist. However, above this threshold, only domination is possible and, in general, local networks become extinct with high probability. Yet, we find that if local networks are launched earlier and for a sufficiently high activity affinity (tendency to engage in more active networks) local networks persist and the international network becomes extinct. We find that in addition to the previously mentioned scenarios, a partial state in which the international network dominates in some countries and local networks dominate or coexist in the remaining is possible as well. Finally, we note that depending on the parameters the final state of the system - whether local networks dominate or become extinct - can be completely unpredictable as it varies randomly between different realizations of the model.

Quite remarkably, a thorough comparison of our model with empirical data from the recent overtake of Facebook indicates that the most probable launch date of Facebook was at the beginning of 2006 and its global launch was in late 2007. Facebook was started in 2004, but opened to the public in 2006, in good agreement with the estimation from our model. Besides, according to Google trend data, 2007 was the year where the global search volume for Facebook started to increase rapidly. And --last but not least-- our best estimation of the model parameters corresponds to the “coinflip” region, which means that the observed overtake of Facebook has only a probability of around 70%. With 30% probability we would be living in a world where each country would have its own successful local network and a network like Facebook would not exist.

Our findings here suggest interesting future research lines. On the one hand, mit remains an interesting task for future research to further increase the precision of the model. This can be done by improving the proxy for the similarity between countries and by adjusting parameters on a country-by-country level. On the other hand, the model can be extended to account for several international networks to investigate their global competition. For a second international network to overcome the first one a certain minimal difference of fitness is needed, which can result from different properties of the networks, like features or functionalities .

:

Finally, random fluctuations of the fitness can be incorporated to describe Darwinian selection in the digital ecosystem.

[1] Kleineberg, K.-K. and Boguna, M. Phys. Rev. X 4, 031046

[2] Kleineberg, K.-K. and Boguna, M. Sci. Rep. 5, 10268

[3] Kleineberg, K.-K. and Boguna, M. Sci. Rep. 6, 25116 (2016)

[4] https://www.youtube.com/watch?v=z3dP3PD7ueA

[5] https://www.youtube.com/watch?v=XkZTxnJd-eI

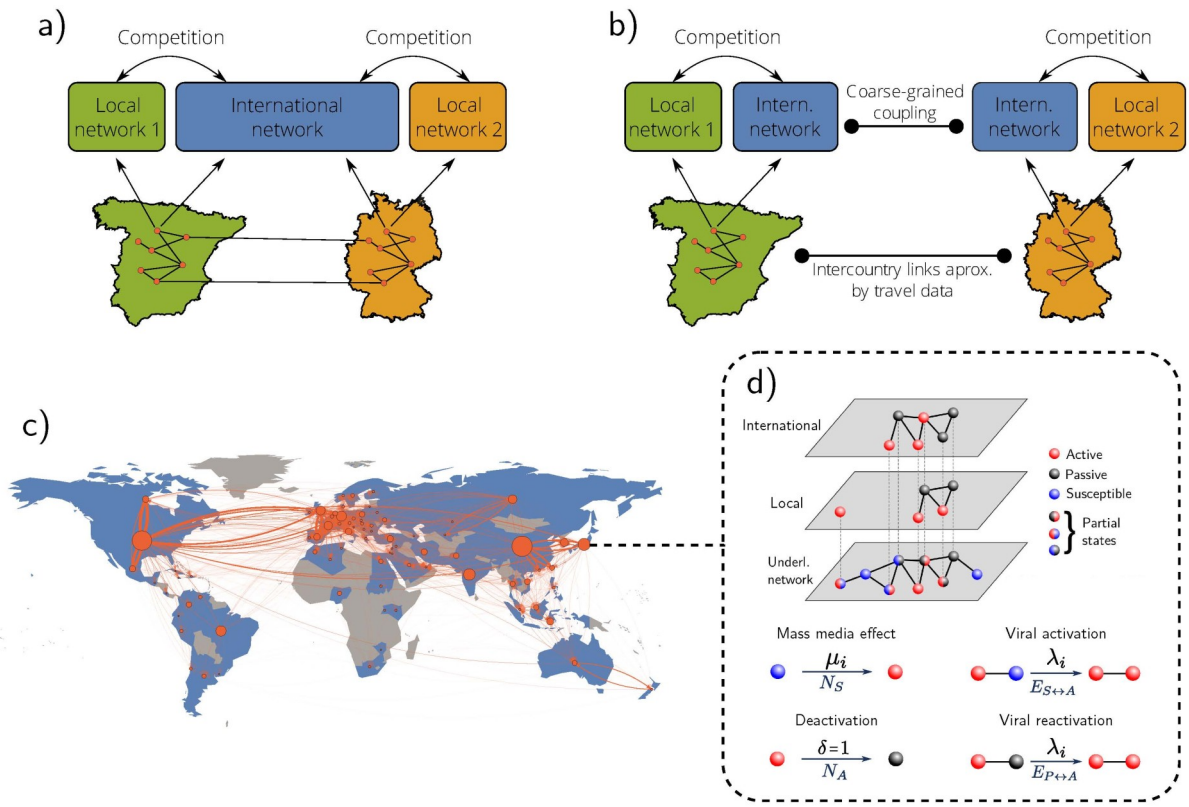
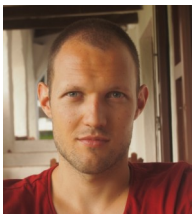


Figure 1 - Network of networks model of the digital world. a) Design of the international network and local networks. b) Sketch of our model using a coarse-grained coupling. c) Visualization of the flight network. The area of the nodes is proportional to the number of Internet users in the respective country with Internet access. The transparency and thickness of the links represents the density of passengers between the involved countries. d) Illustration of the competition between the international and local network within one country

Kaj-Kolja Kleineberg
kkl@ffn.ub.edu
ESR iSocial Fellow
Universitat de Barcelona, Spain



Users Key Locations in Online Social Networks: Identification and Applications

Key locations can be used to characterize user's behavior both in terms of mobility and sentiment

Cities around the world have the need to frequently monitor the commuting patterns of their inhabitants in order to maintain and improve the quality of their transportation systems. With the advent of social media and Internet-enabled devices, users share useful information online, which can be used to identify these patterns. Information about leisure activities, work and other habits are often accompanied with geographical metadata. This flow of information is continuously streamed as live data and, in most cases, is accurate enough to provide information about the exact location of the user at a given time.

In our previews work we present an approach, which analyzes a user's geo-tagged Twitter activity traces to identify her key locations; namely her Home, Work and Leisure areas. With this research, we use the results produced by the proposed approach and examine a number of applications of our method. In specific, we show that user's Key locations can be used to characterize her behavior both in terms of mobility and sentiment. Our results suggest that users tend to live and spend their free time in close proximity to their Work location during weekdays. During weekends, users leisure travel distance increases to locations further away from their Home. Additionally, we observe that the formation of a user's social network is strongly affected by her Home location, with a large number of her connections living in close proximity. Also, the user's stronger connections, as defined by reciprocity, are not only in close proximity but also in areas with similar economic status.

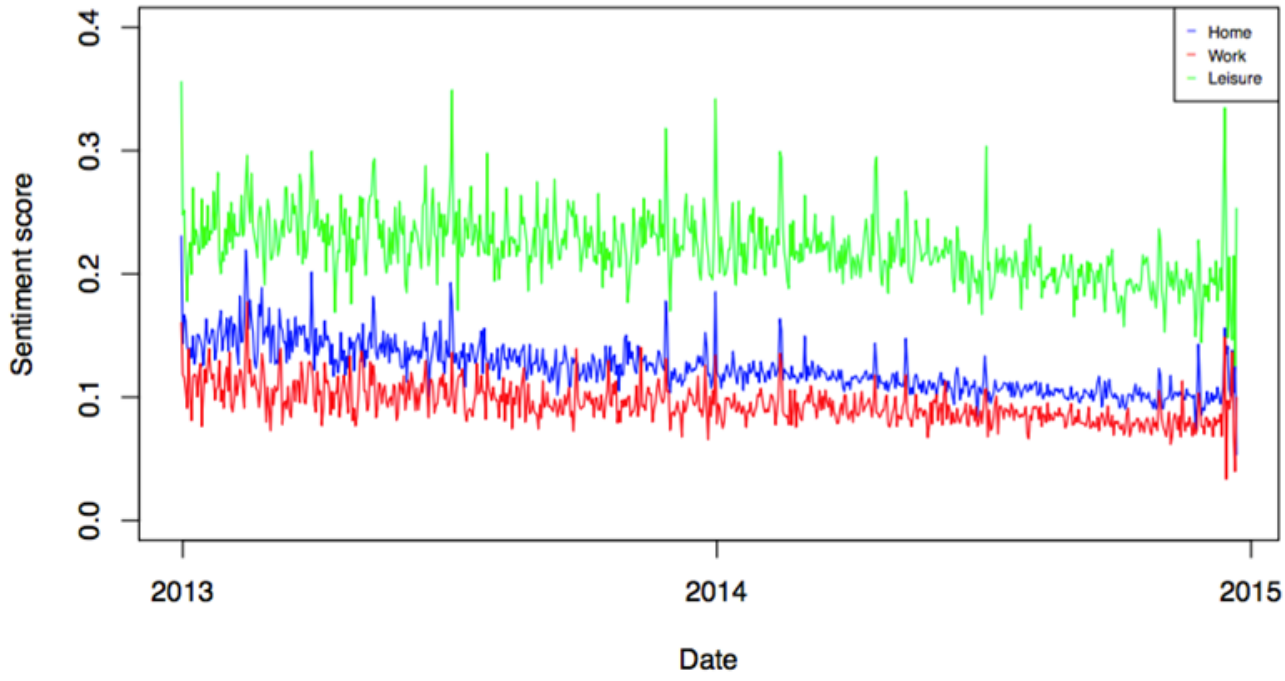
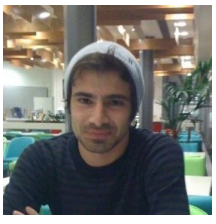


Fig. 1 - **Sentiment per calendar day for the Tweets published from *Home, Work* and *Leisure* areas**

Another part of our research examines the influence of a user's key locations to the sentiment of the text that she publishes in Twitter. Sentiment is commonly used to measure the emotions of user's natural language. It can show the reaction of users to several events or their emotional state during a conversation. Combined with location information it can show how different geographical areas react to specific events or express themselves during their everyday online interactions. Our findings show that users tend to be far more positive in their tweeting behavior when tweeting from leisure locations rather than tweeting from their Home or Work.

Hariton Efstathiades
h.efstathiades@cs.ucy.ac.cy
ESR iSocial Fellow
University of Cyprus (UCY), Cyprus



Building a Spam Free Social Network

As Online Social Networks (OSNs) continue to grow in popularity, a spam marketplace has emerged that includes services selling fraudulent accounts, as well as a nucleus of spammers who propagate large-scale spam campaigns. While OSNs spam has gathered a great deal of attention in the past years, researchers have developed approaches to detect spam such as URL blacklisting, spam traps and even crowdsourcing has been used for manual classification. Although previous approaches have shown the effectiveness of using statistical learning to detect spam, the existing schemes often assume labeled training data. Consequently, such supervised schemes fail to detect new spam content when the spammers change their strategy.

Our work offers unsupervised and massively parallel spam filtering framework that perfectly fits DOSNs requirements

Differently from existing works, in iSocial we have developed an unsupervised framework for spam detection called DLSAS (Distributed Large-Scale Anti-Spam for Social Networks). Our framework employs unsupervised machine learning algorithm, hence it diminishes the need of any prior knowledge of spam patterns. Furthermore, our framework is a fully decentralized and adaptive system that exploits fully decentralized learning and cooperative approaches not only to preserve users privacy, but also to increase the system reliability and to make it resilient to mono-failure.

Decentralized Online Social Networks (DOSNs) have been proposed recently as an alternative for the current OSNs. DOSNs operate as distributed information management platforms on top of Peer-to-Peer network. The main objectives behind decentralization are to preserve users privacy in both shared content and communication, and also to provide complete freedom from any form of censorship or profiling. Although the DOSNs paradigm presents promising ways for preserving users privacy, it creates even more challenges when it comes to the network vulnerability to spam. Indeed, in the absence of a central management entity, designing mechanisms to control propagated or shared content in a DOSN brings up several challenges. First, malicious nodes (i.e., spammers) are integrated in the system, such that they are going to spread the spam while pretending to be good or legitimate users. Second, nodes are restricted by the limited knowledge they have about the network hence they have to collaborative in order to build consensus about the whole network. Third, distributed machine learning algorithms cannot be directly applied in such environment due to data bias of the limited knowledge that every node has.

DLSAS: Distributed Large Scale Anti-Spam

- No a priori knowledge is required.
- Hybrid features:
  - Content-based features:
    - Calculating ratio of Similarity, spam words, URLs, Hashtags, and Mentions.
  - Network-based features:
    - IN to OUT degree
- Gossip Learning: adopt ensemble learning using gossip-based peer sampling.
- Validation overlays: detecting misbehaving nodes and disconnecting them.

Figure 1 - The proposed framework for spam detection

To address the above challenges, we have developed our framework by building three different components as shown in Figure1. The first component is the unsupervised clustering, where every node performs the clustering algorithm using its own local data. Afterwards, in order to construct an unbiased view of the clustering results of the whole network, nodes have to contact a uniform random sample from the network. Therefore, the second component in our framework creates a network overlay that connects every node in the social network with a random set of nodes. The randomness here is required in order to guarantee that every node is connected to a diverse mixture of experts having learning models come from different sources of information. However, as aforementioned the network does not only contain legitimate users, it contains malicious nodes as well. Those malicious nodes can report misleading clustering results to hide their existence. Consequently, we build the third component, which is the validation overlay. The validation overlay is build on-the-fly by randomly selecting nodes that are responsible for collecting data produced by specific nodes and reproduce the clustering results using the collected data. Then, any falsely results can be detected by comparing the reported results from the nodes with the results generated by the validation overlay.

The experimental results using Twitter dataset confirms the ability of our framework to detect spammers with 60% accuracy and less than 3% false positive rate as shown in Figure2.

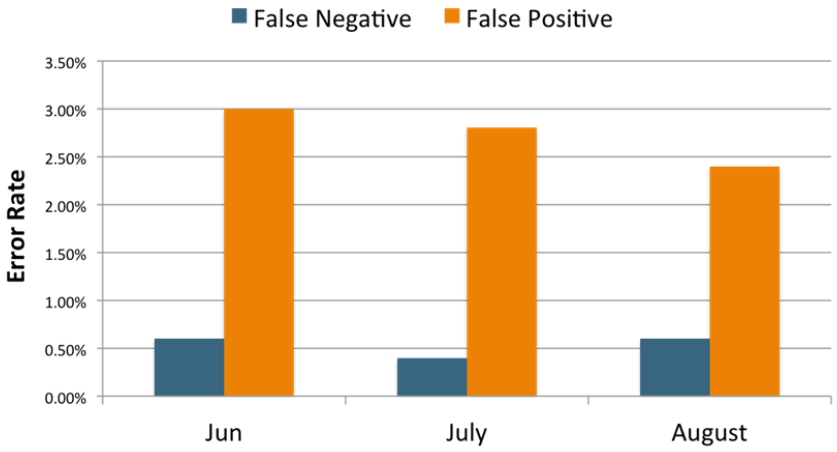


Figure 2 - The results achieved by DLSAS using Twitter dataset.

Amira Soliman
aaeh@kth.se
ESR iSocial Fellow
Royal Institute of Technology (KTH), Sweden

Large Scale Topic Detection using Node-Cut Partitioning on Dense Weighted Graphs

Topic Detection in Text (TDT) is the problem of automatic determination of mutually related documents in a given corpus of text. Extracting topics in online social networks (OSNs) is a challenging problem due to fast speed of publication and large number of noisy messages that respectively affect scalability and quality of proposed solutions. We developed an algorithm to overcome those limitations by combining *Dimensionality Reduction (DR)* with graph analytics. We use a DR method called *Random Indexing (RI)* and a partitioning algorithm inspired from a method called JaBeJa-VC. The solution is composed of three steps:

1. **Creating a Complex Graph:** We use graphs for mathematical representation of the documents. Figure 1 shows our graph construction protocol on a sample of 4 documents in 3 steps. First, every document *D*1, *D*2, ... is transferred into a small fixed length vector *V*1, *V*2, ... using RI (1.1). Second, a *Document Graph (DG)* is created from the correlation matrix of each vector (1.2). Third, DGs are aggregated into a highly dense and weighted graph called General Graph (GG) (1.3)

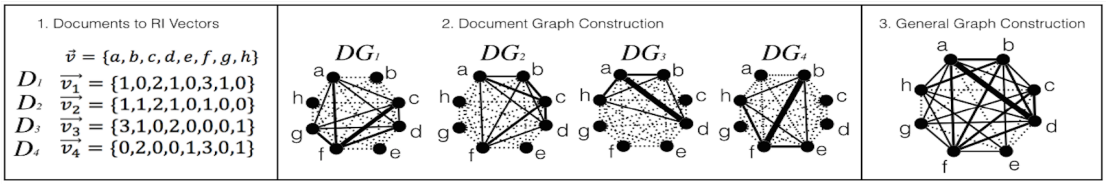


Figure 1 - Graph construction protocol. Shows the General graph (3) resulted from aggregation of Documents Graphs (2) that represent the correlation matrix of their corresponding Random Indexing vectors (1).

2. **Extracting Topics using Partitioning: The** goal of partitioning is to extract weighted dense sub-graphs that represent topics in the GG. Our algorithm presents partitions as colors over the edges of the graph and extracts those partitions by locally accumulating different colors in weighted dense areas of the graph. The algorithm proceeds by maximizing a local utility function in each vertex that results in a global minimization of the node-cut. Each node may belong to multiple partitions hence, node-cut equals to the ratio of the edge-weight of its largest partition over the total edge-weight of the node. In particular, each vertex finds the color of the partition that it belongs to and tries to maximize the volume of that color in its vicinity. Figure 2 depicts a sample partitioning over a small graph. The algorithm starts with a random initialization round followed by multiple iterations until a termination criterion is satisfied. During iterations, every two vertices (e.g., those with a red circle around) try to exchange color of their edges (e.g., e and e') following the mentioned optimization scheme.

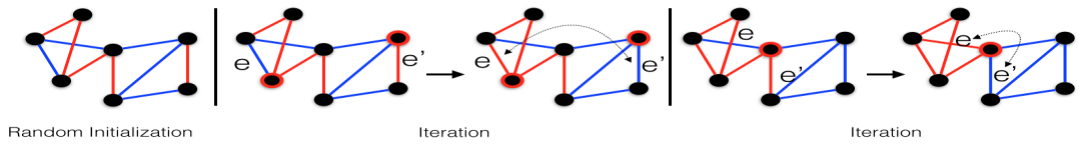


Figure 2 - Sample partitioning begins by random initialization with k=2 and continues in two iterations. Dur-

3. **Assigning Documents to Topics:** topics are assigned to documents relative to the overlap of their corresponding DG to the partitioned GG. We run 3 experiments on datasets from Twitter in 2009. Each dataset contains a set of Tweets with different number of trending topics. Following table shows the distribution of the topics in different experiments. The datasets contain 8*K*, 25*K and* 135*K* documents respectively. Since we don't know the number of partitions in advanced, we first extract a large number of partitions with high precision and low recall and then, merge them using a sampling approach to improve the recall.

Topic	1	2	3	4	5	6	7	8	9	10
Exp 1	3688	1615	1262	1056	705	-	-	-	-	-
Exp 2	6841	6765	4511	2440	2195	1385	1177	360	-	-
Exp 3	85235	24777	6841	6760	4345	2440	2179	1385	1177	360

Table 1 - Topic distributions in different experiments (Exp1, Exp2 and Exp3)

The goal of the algorithm is to represent partitions as colors of the edges and extract those partitions by locally accumulating different colors in weighted dense areas of the graph

Figure 3 shows the F-Score results of different experiments. Each chart corresponds to 100 runs over the corresponding dataset. We used 2% sampling on exp1 and exp2 and 20% on the exp3. As we can see the algorithm achieves highly significant F-Score on the first two experiments with more than 60% average F-Score on all topics. The only drawback is shown on the last 3 topics of the dataset with 10 topics. This is an expected result that emerges due to skewness in the distribution of the size of the topics (as shown in the table above). Those drawbacks are expected to be resolved using a higher number of partitions *K*.

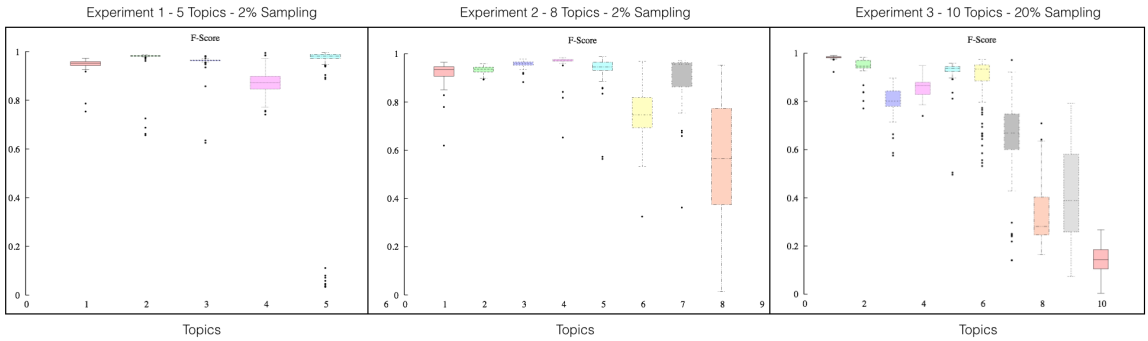


Figure 3 - F-Score values on different experiments.

Kambiz Ghoorchian
kambizgh@kth.se
ESR iSocial Fellow
Royal Institute of Technology (KTH), Sweden

Individualism and Collectivism in Social Dynamics

Opinion distribution in social networks is often modeled with some variation of voter model, where each member of the network adopts an opinion of randomly selected fist-level connection. Many constructions in this context is already thoroughly investigated, including opinion adoption modalities (conformist or contrarian), interaction speed variations, ”stubbornness” of social networks members etc.

Considering the dynamics of voter process on complex networks we constructed the stochastic voter model with distributed flip rates and spontaneous opinion fluctuations in the Langevin approach. The opinion of each agent at each time step can be changed in two possible ways: the agent adopts a state of randomly chosen social contact with some probability or spontaneously enter a random state independently of others nodes. The main point of interest to us is how strong is the impact of rate parameters and stochastic fluctuations of the personal attitudes on contact process, in particular on the average time to reach consensus, and how these parameters are interrelated.

We focused on the case when all population divided into two groups with fast and slow flip rates parameters and equal probability of spontaneous flips. Besides, we consider nodes placed in a mean-field random network with homogeneity in degree distribution.

How to model the emergence of social phenomena like leadership and herding behavior in large populations? What the role of feedback mechanism in collective dynamics? How strong is the impact of stochastic fluctuations of the personal attitudes on contact process?

We have found and analytically proved the presence of a phase transition between two clearly distinct regimes: a purely diffusive phase, as in the standard voter models, and a herding phase (see Fig.1) where a fraction of the agents self-organizes and leads the global opinion of the whole population.

If the fast flip rate is great enough compared to the slow flip rate and the size of the fast group and spontaneous state changing probability is small, the global opinion has sharp discontinuities followed by changes in growing tendency. By contrast, in other cases consensus solutions disappears from fast autonomous dynamics. Instead the fast group dynamics are fluctuating around the opinion of slow group.

Also we considered the feedback mechanism between fast and slow subgroups. When the fast group is in consensus, it attracts slow dynamics to its position in exponential decay. During this process, slow agents also influence the fasts with their mean opinion shifting them from consensus states. For evaluating this mechanism we calculated the transfer entropy between two groups fast and slow agents and proved the present of phase transition.

Thus, we have modeled the emergence of spontaneous leadership and herding behavior in large populations and described the feedback mechanism causes the appearance of these phenomena.

Bibliography

[1] Gardiner, C. W. et al. (1985) Handbook of stochastic methods, volume 3, Springer Berlin.

[2] Boguñá, M., Castellano, C., and Pastor-Satorras, R. Mar 2009 Phys. Rev. E 79, 036110.

[3] Mosquera-Doñate, G. and Boguñá, M. May 2015 Phys. Rev. E 91, 052804. [4] Schreiber, T. Jul 2000 Phys. Rev. Lett. 85, 461–464.

Liudmila Rozanova
l.temereva@ffn.ub.edu
ER iSocial Fellow
Universitat de Barcelona, Spain

Social Network archaeology: Revealing past tendencies from contemporary data

The structure and dynamics of Online Social Networks (OSNs) have been thoroughly studied. Existing methods allow to measure metrics that indicate how sparse or dense is a network what is the role of central hubs and how these properties evolve over time. Unfortunately the majority of network properties that reveal this valuable information require computation time that is prohibitive given the enormous size of modern social graph such as Twitter. As of today Twitter's social graph contains 600 million nodes and tens of billions of edges. Any algorithm with computation or memory requirement above linear makes these estimation a very tedious task.

One network property that is easy to evaluate is the average degree. This is the average number of in-going and out-going edges of the nodes for the complete network. In the past various mathematical models have been suggested for the evolution of this property. Researchers have proposed that a network is going through various growing phases according to the function that optimally fits the evolution of the average degree. This functional can be either superpolynomial, logarithmic or linear. Studying this behavior in Twitter entails many difficulties. The first is that Twitter does not provide direct access to historic data and does not reveal the edge creation time of the social graph. To overcome this we applied a heuristic that approximates the edge creation time. Subsequently, we evaluated these models on two subsets of Twitter's social graph. The first was acquired at the end of 2009 and contains dense subset of Twitter at that time and the second is a smaller dataset acquired at 2015.

The modeling of the average degree on these two datasets revealed an unexpected finding. On both datasets it is clear that Twitter did not undergo a smooth growing phase. In contrast the growth fluctuates intensely between superpolynomial and logarithmic periods. This fluctuation lasted almost 3 year (2006-2009). After August 2009 and until today Twitter seems to undergo a smooth growth of linear increase. We have also identified real events like, technology upgrades and publicity issues that coincide with these fluctuating periods although additional work is required in order to make accurate associations. It is important to note that the second dataset was acquired more than 5 years after the end of the 'fluctuating period', yet it is as accurate as the first in depicting these tendencies. Therefore we have concluded that fitting the average degree into various models is a very efficient (O(n)) method to accurately identify varying growth periods on Social Networks. Moreover we believe that even when OSNs are stringy when it comes to accessing the social graph, the application of sophisticated methods can reveal intrinsic features of their structure, evolution and history.

Figure 1. Fluctations of growth of average degree on Twitter's early period

Despoina Antonakaki
despoina@ics.forth.gr
ESR iSocial Fellow
Foundation for Research and Technology-Hellas (FORTH), Greece

Project Coordinator:
Šarūnas Girdzijauskas
Royal Institute of Technology , Stockholm, Sweden

Newsletter Content Editor:
Maria Poveda
Laboratory for Internet Computing (LINC)
University of Cyprus (UCY)

For more details contact: info@isocial-itn.eu

The project is funded by the European Commission under the Marie Curie Initial Training Network (ITN)

Find us on facebook

https://www.facebook.com/ISocialMarieCurieITN