iSocial

Issue 2, May 2015

iSocial: Decentralized Online Social Networks

1-4 June 2014: "Distributed Technologies for Social Networks" Summer School, Stockholm, Sweden

Inside this Issue

Trending Topics on Decentralized Online Social Networks

Decentralized App Store for a B2B Social Network

The widespread of mobile devices and social networks have favored the usage of App Stores as a tool for the massive distribution of software components. Our project focuses on the development of a Distributed App Store for B2B Social Networks. This App Store deals with the basic operations, plus others derived from the B2B requirements of the platform.

See Page 4

With the Assistance of Social Dynamics. A Decentralized Topology Construction Protocol for a Decentralized Online Social Network

Designing a Decentralized Online Social Network (DOSN) on top of P2P architecture presents a series of challenges. Decentralized topology construction protocols play a prominent role on the solution of the above challenges such as data storage, data availability etc. This project focuses on the design of a decentralized P2P topology suitable for a DOSN system and on how the social dynamics can be integrated without changing the peers' uniform distribution and the overlay network.

See Page 5

Location Based Access Control for Video Streaming

Online Social Networks constitute a popular communication technology for which Video Distribution is a fundamental building block. While social interactions are moved on the Internet, privacy is perceived by users as an increasingly important requirement. The same type of communication paradigm gets applied within enterprises, where the exchanged data might contain highly confidential information, and security becomes even more relevant.

See Page 6

Fault Tolerance for Distributed Stream Processing Engines

Distributed stream processing engines (DSPEs) handle continuous stream of data and guarantee real-time processing. To handle the input stream efficiently, DSPEs require clusters of thousands of machines. Also, DSPEs operate for indefinite period of time. The use of larger number of machines and the need of operating for indefinite period increase the likelihood of failures. In our work, we study the fault tolerance problem and propose various solutions to handle such failures for DSPEs.

See Page 7

Fully Distributed Risk Assessment based on Detecting User Anomalous Behavior in DOSN

As the number of users in Decentralized Online Social Network (DOSN) increased, the DOSN grows in size and complexity. Therefore, this can attract a variety of highly damaging attacks. Researchers have observed attackers forwarding spam and malware on DOSNs. Since attackers and risky users have weird behavior pattern in the network, our goal is to analyze the behavior of users in DOSN by identifying misbehaviors to detect risky users.

See Page 8

Preserving User Privacy in Shared Photos

The widespread adoption of online social networks (OSN) has raised many concerns regarding users' privacy. The social networking services spend effort on preserving users' privacy by designing and implementing mechanisms that allow users to define their preferred access control policy. However, the existing access control mechanisms consider the uploader as the owner of the data and grant him full rights, but they do not grant any rights to the people that are related to this data.

See Page 10

Protecting Users' Data and Privacy on the Social Web – Does it require new inventions?

It is not a new statement that users' personal data in the realms of the social web needs to be better protected. Marketers, data brokers, research institutions, governmental units, burglars, or simple curious friends, are examples of entities that exploit the richness of the information we unwillingly, or willingly, unconsciously, or consciously, make at their disposal.

See Page 11

Delving into Twitter's Social Network

Twitter is one of the most popular social networks with an inherent simple design and a huge user base. We can model users and following relationships in Twitter as nodes and edges respectively. This is called the social graph and it can help us to visualize the structure of Twitter and reveal interesting patterns.

See Page 13

Ecological Theory of the Digital World

The success of Web 2.0 has changed the way humans interact on a worldwide scale through online social networks. The success of these networks is based on their capacity to engage users whose time is a limited resource. The digital world thus forms a complex ecosystem of interacting networks. Analogously to the case of population dynamics, the question: "Why do we observe a moderate number of coexisting digital services?" arises and remains unanswered.

See Page 15

Identification of Key Locations based on Interactions in Online Social Networks

Ubiquitous Internet connectivity enables users to update their Online Social Network profile from any location and at any point in time. These, often geo-tagged, data can be used to provide valuable information to closely located users, both in real time and in aggregated form. However, despite the fact that users publish geo-tagged information, only a small number (~30%) publicly reveal explicit accurate information about their location in their profiles with granularity higher than city level. In our research we focus on designing and implementing an effective methodology for identifying a user's key locations, namely her home, work and leisure places.

See Page 16

Building a Smarter Social Network

We're working on building a smarter future for social networking at large, and big part of our vision is secure and privacy preserving environment. Therefore, we propose Smart Social Networks (SSNs) with the potential to engage more effectively and actively with users to bring the peak benefit of big-data. Decentralized Social Networks (DOSNs) are proposed with the potential to reconsider the well-being in providing services closer to users' needs and expectations.

See Page 17

Distributed Semi-Supervised Multiple Disambiguation

An ambiguous mention is a word or a phrase that can be used to name multiple different entities. Disambiguation is the task of categorizing a set of documents containing a specific ambiguous mention into a set of groups such that all the mentions in each group refer to the same entity. An algorithm was recently proposed in our group that overcomes scalability issues of the current solutions by applying a distributed processing approach.

See Page 19

Network Modeling and Overlay Design for DOSNs

The specification of the overlay network, the topology of connections between peers that are used to manage the services provided by the platform in a decentralized manner is an important aspect of the design of a Decentralized Online Social Network. The online social network layer should be advantageously used in this design. Network modeling can offer useful insights for both the social and the overlay network layers.

See Page 21

Decentralized App Store for a B2B Social Network

The widespread of mobile devices and social networks have favored the usage of App Stores as a tool for the massive distribution of software components. The benefits of these tools are twofold. First, they allow developers to access a very large user database with smaller effort. Second, they enable providers to greatly enrich the users experience seamlessly inside the social network.

The development of App Stores is complex and includes challenges such as security (authentication and authorization), search, distribution, storage, update propagation, social capabilities (rating, comments), etc. These challenges are augmented when economic transactions are introduced, and their complexity increases even more when they are performed in a distributed fashion.

Our project focuses on the development of a Distributed App Store for B2B Social Networks. This App Store deals with the basic operations, plus others derived from the B2B requirements of the platform.

Our project focuses on the development of a Distributed App Store for B2B Social Networks. This App Store deals with the basic operations, plus others derived from the B2B requirements of the platform. The B2B oriented App Store must provide support for secure economic transactions, different payment models (single payment, subscription, pay-per-use) and revenue sharing of income between different actors.

So far all developments on App Stores have been made following a centralized model. To the best of our knowledge, models for decentralized social networks are made on simplified App Stores not accounting for the rich complexity of the problem.

In our project, we propose a semi-decentralized model where some operations (App distribution, update propagation) are performed in a peer-to-peer fashion by clients, while other ones (payment processing, revenue sharing) are maintained by the servers.

In our model, the servers are the initial point of distribution of apps and updates. Once apps are first delivered by the Store to an initial set of clients, these nodes start to distribute the software to the rest of clients in a peer-to-peer fashion, while the servers act as backup in case the app is not available in any other node. A set of use cases will provide information to decide on the best possible network topology for the caching and distribution of apps.

This model leverages some of the complex operations of the App Store to the network, while keeping the operations that require a central authority centralized. This approach reduces the dependency on centralized services and is intended as a stepping stone on the way to a fully distributed, peer-to-peer App Store to empower distributed social networks.

Andrés García García
andresg@il.ibm.com
ER iSocial Fellow
IBM Research Lab Haifa, Israel

With the Assistance of Social Dynamics. A Decentralized Topology Construction Protocol for a Decentralized Online Social Network

Designing a Decentralized Online Social Network (DOSN) on top of P2P architecture presents a series of challenges, such as data storage, data availability, data consistency, network topology, etc. Decentralized topology construction protocols play a prominent role on the solution of the above challenges, as well as the implementation of a DOSN. These protocols have received a large amount of attention due to their ability to distribute the peers of the physical network (IP) uniformly and progressively organize peers into a logical structure, called overlay network. The overlay network is responsible to support the routing, search and key-value functionality, as well as maintain the scalability (number of peers in the network) and robustness (resilient against high churn) of the DOSN system.

However, integrating the social dynamics into the design of a decentralized P2P topology is a challenging task, as the uniform distribution of the peers needs to be maintained. Current implementations of DOSN systems assume no correlation between the social connections of the OSN and the decentralized topology construction. As such, the routing, search and key-value functionality of the DOSN relies on the complexity that the overlay network provides, which is linearly increasing based on the number of social friends that each social user maintains. For instance, when a social user generates a status update and he wants to inform all of his friends about the update, he needs to lookup into the overlay network for each one of his friends in order to identify the friends' exact position, providing though a linear complexity for the lookup functionality based on the degree of the social user. Therefore, status update functionality requires thousands of lookups and can take hundreds of seconds to generate.

My goal is to create a topology construction protocol where the links between the peers into the overlay network be selected in a manner that exploits the social friendship.

My project focuses on the design of a decentralized P2P topology suitable for a DOSN system and on how the social dynamics can be integrated without changing the peers' uniform distribution and the overlay network. My goal is to create a topology construction protocol where the links between the peers into the overlay network be selected in a manner that exploits the social friendship. Based on this logical structure, when a peer into the overlay network looks up for one of his social friends, the possibility to be directly connected into the overlay network will be increased, and as such our proposed DOSN will not suffer from high traffic overhead.

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

Location Based Access Control for Video Streaming

Online Social Networks constitute a popular communication technology for which Video Distribution is a fundamental building block. While social interactions are moved on the Internet, privacy is perceived by users as an increasingly important requirement. The same type of communication paradigm gets applied within enterprises, where the exchanged data might contain highly confidential information, and security becomes even more relevant.

The goal of this research project is to develop a pragmatic architecture for Location Based Access Control in P2P Video Distribution Systems.

Besides the classic Role Based Access Control, in which users are grouped in classes and assigned with common sets of permissions, the enterprise scenario requires additional kinds of restrictions. Location Based Access Control allows to define and enforce policies based on the physical location of the user, preventing for instance the visualization of confidential material outside the office.

Within this context, I am conducting a research project in collaboration with the University of Insubria, with the goal of developing a pragmatic architecture for Location Based Access Control in P2P Video Distribution Systems.

In the designed protocol the location is defined with respect to Wireless Access Points. The physical locations claimed by a node can be verified by other nodes as the knowledge of a shared secret, which can be known only within authorized positions. Certification of claimed positions is achieved by means of a public key infrastructure. The P2P infrastructure is responsible for the enforcement of the policies, and delivers the video content only to nodes which can prove to be located in an authorized area.

Giovanni Simoni
giovanni.simoni@peerialism.com
ESR iSocial Fellow
Peerialism AB, Sweden

Fault Tolerance for Distributed Stream Processing Engines

Distributed stream processing engines (DSPEs) handle continuous stream of data and guarantee real-time processing. DSPEs target applications related to continuous computation, online machine learning and real-time query processing. For example, set of applications includes fraud detection, network monitoring, anomaly detection, financial transactions and others. Such frameworks operate on high volumes of data that arrives at a rapid rate from various sensors such as web traffic, call data records and stock markets.

> ***The purpose of our work is to study the fault tolerance problem and to provide a set of solutions for various stream applications. We study various reliability models for DSPEs and propose a set of fault tolerance techniques for streaming frameworks.***

DSPEs deal with high volumes of data by applying lightweight operations on the real-time and continuous streams. To handle input streams efficiently, DSPEs need clusters of thousands of machines. The use of a larger number of machines for deployment of a DSPE increases the probability of failures that includes network, software and hardware failures. These failures can produce erroneous results. However, various streaming applications are critical and intolerant to failures, e.g., financial transactions and radar systems. Such applications cannot afford loss of information. To handle different failure scenarios, DSPEs require being fault tolerant.

Most of the practical deployments of DSPEs at companies like Yahoo and Google use thousands of machines. The use of a larger number of machines increases the likelihood of failure of a single machine inside a cluster at any given moment. Furthermore, streaming applications run for indefinite periods, making it almost impossible to avoid failures. Failure in DSPEs can cause an application to stop or produce erroneous results. As a consequence, along with scalability and low latency requirements, DSPEs also require being highly fault tolerant.

The purpose of our work is to study the fault tolerance problem and to provide a set of solutions for various stream applications. For example, financial applications cannot afford any loss of data during transactions, whereas, filter operators can operate efficiently even with loss of some information. In our work, we study various reliability models for DSPEs and propose a set of fault tolerance techniques for streaming frameworks. The final product will be a detailed study of fault tolerance techniques and a set of techniques that can be applied to DSPEs.

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

Fully Distributed Risk Assessment based on Detecting User Anomalous Behavior in DOSN

As the number of users in Decentralized Online Social Network (DOSN) increased, the DOSN grows in size and complexity. Therefore, this can attract a variety of highly damaging attacks. Researchers have observed attackers forwarding spam and malware on DOSNs. Since attackers and risky users have weird behavior pattern in the network, our goal is to analyze the behavior of users (interactions or activity patterns) in DOSN by identifying misbehaviors to detect risky users. The basic idea is the more the user behavior diverges from "normal behavior", more the user is risky.

> ***Our goal is to solve risk assessment problem by proposing a risk assessment model over fully distributed data.***
> ***A key problem in our risk assessment model is how to minimize the communication overhead and energy consumption in the network.***

Risk assessment over fully distributed data where each user has a single feature vector and his/her interactions and personal information cannot be moved to a central server or to other users due to privacy issues, poses an important problem in this social network. All users have some features, like number of friends, posts, likes, comments, average number of mutual friends, etc. where these feature values never leave the computer of a user in a raw form due to the privacy considerations and furthermore their feature values can also change over time. In addition, this social network is dynamic (users can join and leave any time) and also is vulnerable to all kinds of attacks.

Solving all these problems for risk assessment by monitoring (analyzing) the activity pattern of users in a distributed way in DOSN is a grand scientific challenge. Our goal is to solve this problem by proposing a risk assessment model over fully distributed data. A key problem in our risk assessment model is how to minimize the communication overhead and energy consumption in the network.

On the other hand, it is essential to process the personal behavioral data records of users locally. But, it is not possible to learn from local models because of the lack of information. On the contrary, there is no possibility to build local learning models and combining them similar to hierarchical models. Because, this open the opportunity for attackers to become a leader and collect the local models for the final prediction. Besides, the communication cost needs to be kept low during our learning process. Working towards this goal, we propose a gossip learning approach involves all users with their local estimation of the learning model and applying an online learning algorithm to improve their estimations and finally converge to a global learning model.

Our gossip learning approach involves multiple local estimations of the learning model. It exchanges these estimations between users over the social network in parallel and updates them. The purpose of the protocol is to disseminate up-to-date information and maintain them without collecting them in the central place, in a dynamic large-scale social network. The basic underlying idea is that all users are equivalent and run the same learning model periodically. Then they exchange their local estimation with each other to update them and finally to converge a global model.

The design of our risk assessment learning model requires specific aspects as the following: the risk model has to be robust and it should maintain a reasonable performance even in extreme failure that a lot of users leave the network or don't respond the messages. Also, all users should be able to assign a risk score to other users and make prediction immediately at any time in a local way. In addition, we avoid attackers to manipulate the learning model during the gossiping process. Finally, the global risk model has to have a low communication cost by decreasing the number of messages and the size of them as well.

In order to evaluate our distributed scheme, we will implement our algorithm in a real Facebook data set. Our goal is to achieve a comparable accuracy compared to a centralized scheme and a significant reduction in communication overhead. Also to preserve the privacy of user's information, avoiding attackers to manipulate the learning model.

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

Preserving User Privacy in Shared Photos

The impressive popularity and widespread adoption of Online Social Networks (OSN) has raised many concerns regarding users' privacy. Unfortunately, most online users are not concerned about their privacy and tend to disclose sensitive information publicly. One of the main reasons this happens, is that users are usually oblivious to the true visibility of the data and information they publish online. The social networking services spend effort on preserving users' privacy by designing and implementing mechanisms that allow users to define their preferred access control policy. This policy specifies which particular user groups should be permitted or restricted from accessing each data object. However, the existing access control mechanisms consider the uploader as the owner of the data and grant him full rights, but they do not grant any rights to the people that are related to this data. As a consequence, users are allowed to specify their desired policy only for self-published data, but not for information that was disclosed by another user.

We design a mechanism such as the privacy settings of each user are enforced according to his particular needs, and cannot be overwritten by those of another user, or by those of the photo uploader.

This problem becomes more evident in the case of shared photos, as the identity of the users depicted in a photo can be possibly revealed, without them being able to prevent it. At first, the uploader is not required to request the permission of depicted users before publishing the photo. Also, users do not have the permission to remove a photo uploaded by others. They can request the removal of such a photo, but the decision for keeping or removing it remains entirely on the uploader, which may not be concerned about privacy and thus, not having the incentive to fulfill their request. Figure 1 illustrates the privacy risk of a user depicted in a shared photo (square node). The green nodes represent the friends of the particular user, while the red nodes represent the OSN users gaining access to the photo because of other depicted users.

This problem can be solved by a mechanism that allows each depicted user to contribute on the specification of the photo's access control policy. However, this mechanism should be able to effectively handle the cases where conflicting permissions are set by the users. In order to avoid conflicting cases, we design a fine-grained mechanism that changes the granularity of access control from the level of a photo to that of users' regions of interest. Specifically, this mechanism identifies orthogonal users' regions of interest in the shared photo, and allows each user to determine his desired rules for specific regions. Thus, the privacy settings of each user are enforced according to his particular needs, and cannot be overwritten by those of another user, or by those of the photo uploader.

This mechanism takes into consideration the identity of the access requesting user. Also, it efficiently determines at real time the regions of the photo that should be permitted or restricted according to the depicted users' policy. It should impose low overhead for not affecting user experience, and also, we consider low complexity and high usability as a requirement, for attracting users into adopting it. Furthermore, such a mechanism should be easily implementable and compatible with the existing access control mechanisms, for being adopted by the current social networking services.

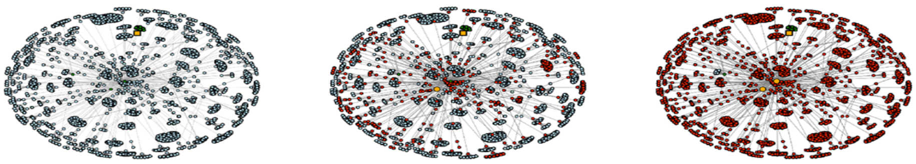


Figure 1: An example that illustrates the privacy risk of a user depicted in a shared photo

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

Protecting Users’ Data and Privacy on the Social Web – Does it require new inventions?

> *“Now, I daresay you noticed, that last time you picked me up, that I was looking rather thoughtful?'*
>
> *`You were a little grave,' said Alice.*
>
> *`Well, just then I was inventing a new way of getting over a gate -- would you like to hear it?'*
>
> *`Very much indeed,' Alice said politely.*
>
> *`I'll tell you how I came to think of it,' said the Knight. `You see, I said to myself, "The only difficulty is with the feet: the head is high enough already." Now, first I put my head on the top of the gate -- then I stand on my head -- then the feet are high enough, you see -- then I'm over, you see.'*
>
> *`Yes, I suppose you'd be over when that was done,' Alice said thoughtfully: `but don't you think it would be rather hard?'*
>
> *`I haven't tried it yet,' the Knight said, gravely: `so I can't tell for certain -- but I'm afraid it would be a little hard.'”*

Our current research work aims at designing solutions that would allow a-posteriori access control for better data security and users’ privacy in online social networks.

Alice, in the eighth chapter of the sequel of “Alice’s Adventures in Wonderland”, had the chance to meet the white knight and learn about his splendid own inventions. After amazing her with an invention of his own to keep the hair from being blown off by the wind, simply based on the fact that “the reason hair falls off is because it hangs down” because “things never fall upwards”, he dazzled her by his more interesting invention on how to get over a gate! Should we adopt the approach of the white knight and invent new ways for protecting users’ data privacy in the social web? I am afraid this is a question only Alice could answer. Nonetheless, our responsibility as researchers is not only to solve new problems, but is also to find better ways for achieving known goals. Do these new ways have to be pure new inventions? Yes, they could be, but they also could be as simple as adapting some old practices to new scenarios. That being said, we leave new inventions to Alice’s white knight, and we explore how an old practice can improve the way we approach users’ privacy.

It is not news that users’ personal data in the realms of the social web need to be better protected. The proliferation of online social networks (OSNs) and of the amounts of interconnected personal data generated and published in the web, makes us exposed, more than we ever could imagine, to whoever is interested in getting some benefit from us. Marketers, data brokers, research institutions, governmental units, burglars, or simple curious friends, are examples of entities that exploit the richness of the information we unwillingly, or willingly, unconsciously, or consciously, make at their disposal.

So far, the protection of users’ personal data in OSNs has been governed by access control models that focus on managing locking and unlocking operations on protected resources based on the security criteria of their owners [1]. This approach, being based on an apriori management style, might be fitting well for controlled environments within which patterns of data generation, of data ownership, and of data access, are well defined and are well known beforehand; however, it demonstrates serious limitations where these cannot be anticipated. Besides, apriori access control stops where incidents happen, as locks become meaningless once data leaks through them. In fact, real life events have confirmed to us that data, sooner or later, inevitably escapes its locks. It follows, at no surprise, that information security and privacy advocates call for new ways for managing users’ data. For instance, authors of [2] raise the awareness of the research body in the privacy field to the need for a paradigm shift that takes personal data security from apriori access control management only to a continuous protection process that is based on transparency and that ensures accountability. Do we then need a new invention for answering this call? It seems not, as our ancestors have, since ancient times established the pillars of what is today known as auditing.

Auditing, “from Latin audītus, a hearing” [3], can be defined as “a thorough examination or evaluation” [3] of some records. It is a well-established field and a necessary practice in finance and accounting. According to historians of accounting, “there is evidence that the government accounting system in China during the Zhao dynasty (1122 – 256 BC) included audits of official departments” [4]. According to the same source, Romans and Greeks, since as early as the fifth and fourth centuries BC, designed systems of dutiful checks and counterchecks to verify the accuracy of their reports. In today’s business, as well as in the general management of our social and legal structures, auditing is a required activity for healthy environments. Indeed, the mere knowledge of the existence of records that hold us accountable if we break the rules, seems to be the basis for our general respect to social, legal, institutional, and other set rules that govern different aspects of our personal and professional lives. As has been dutifully argued in [5], in the presence of transparency (defined as “established norms for open communication”) it is much less probable to see “destructive individuals” get started with “clandestine acts”. Moreover, as “accountability refers to the presence of limits to individual discretion”, the risk of destructive behavior is further limited. This clearly suggests that transparency and accountability make compliance to rules more likely than violation compared to when these two elements do not make part of a system’s climate.

With the development of information technologies and with their adoption in and invasion to different domains and life applications, these concepts of transparency, accountability, and auditing have not taken long to be considered for information systems as well. As such, a body of research work addressed the study, design, and implementation of a-posteriori approaches to data security management for some specialty systems [6]. However, this a-posteriori paradigm has not been investigated for OSNs yet; though these constitute one of the scenarios that best exemplify the limits of a-priori only solutions to data security management. From this standing point, our current research work aims at designing solutions that would allow a-posteriori access control for better data security and users’ privacy in online social networks. The main challenge with this is in the distribution of reporting and auditing tasks and in the design of techniques that would enforce them without falling in the pitfall of a central entity owning and controlling all the environment. Our approach to address this is that users are not to be viewed as data generators and OSN services’ customers only, but they are also a promising power-force that can be recruited, in a collaborative effort, as watchdogs, whistle-blowers, official reporters, auditors, and accountability managers.

With a strong belief in the need for transparency and accountability in the management of data in online social networks, we started our work by designing a framework for report and audit based data sharing and by instantiating it for a decentralized social network model. Our aim is not to make new inventions of our own, but to help establish standards and practices that have proved efficiency and effectiveness across several critical domains within the realms of future online social computing.

References:

[1] E. Ferrari, Access Control in Data Management Systems, ser. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2010.

[2] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, “Information accountability,” Communications of the ACM, vol. 51, no. 6, pp. 82–87, 2008.

[3] The Free Dictionary, “Auditing”. Accessed in March, 2015. [Online]. Available: http://www.thefreedictionary.com/auditing

[4] M. L.Pava, “Encyclopedia Britannica-Auditing, Accounting”. [Online]. Available: http://www.britannica.com/EBchecked/topic/42575/auditing

[5] D. Day, The Oxford Handbook of Leadership and Organizations. Oxford University Press, 2014.

[6] K. Padayachee and J. H. Eloff, “Adapting usage control as a deterrent to address the inadequacies of access controls,” computers & security, vol. 28, no. 7, pp. 536–544, 2009.

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

Delving into Twitter's Social Network

Twitter is one of the most popular social networks with an inherent simple design and a huge user base. It currently has 700 million users, it is very mobile-friendly and it has an extensive and well-documented programming interface. We can model users and following relationships in Twitter as nodes and edges respectively. This is called the social graph and it can help us to visualize the structure of Twitter and reveal interesting patterns.

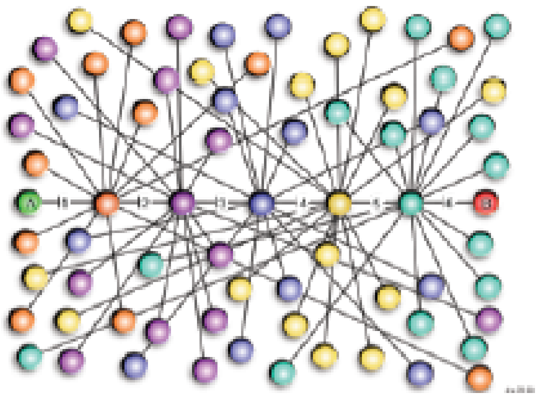
The high connectivity and scale-free nature of social networks have been already studied from the 50s where the famous "six degrees of separation" paradigm originates. According to this, six hops in a social network are enough to go from any node to another.

The availability of Twitter's network combined with the progress on graph theory and computation over the last years allows us to investigate even more exciting properties.

Furthermore, the availability of Twitter's network combined with the progress on graph theory and computation over the last years allows us to investigate even more exciting properties. For example algorithms can perform community detection and locate the virtual neighborhoods and affiliations of each user.

Subsequently a recommendation system can utilize this information and suggest users that have similar interests/ideas. Another example is the quantification of the influence of a user within a virtual community, or else the measure of the impact (positive or negative) of an individual's tweets. Needless to say, the number of a user's followers is not the most contributing factor to this user's influence in a community.

Special metrics (for example: betweenness centrality), in a modified social graph that takes retweets rather than followings as edges, is a more indicative metric of a user's influence.



http://en.wikipedia.org/wiki/Six_degrees_of_separation

Pictures from Marvel, S. A.; Martin, T.; Doering, C. R.; Lusseau, D. & Newman, M. E. J. (2013), 'The small-world effect is a modern phenomenon.', CoRR abs/1310.2636 .

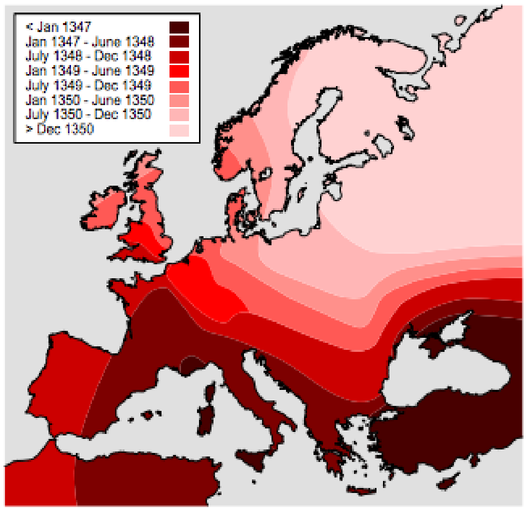
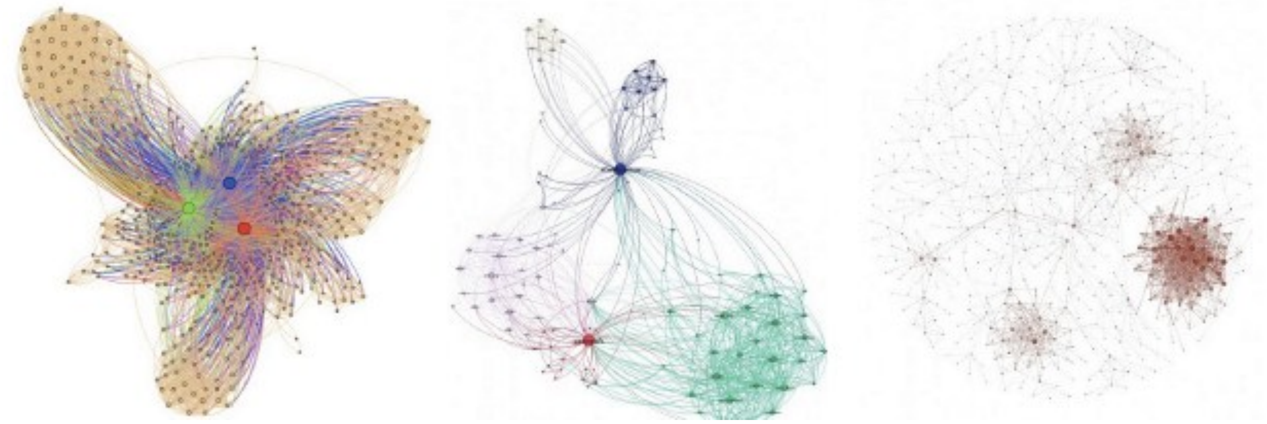


FIG. 1: The spread of the Black Death across Europe in the 14th century, after Sherman and Salisbury [18]. Observe that the disease advanced as a wave of infection across the continent at a more or less constant speed for over three years.

With these techniques, we can locate influential people in an online community and we can measure (and sometimes even predict) the collective reaction of an action or event. Given the increase of popularity and prevalence of online social networks, we can expect that these metrics will be taken into consideration from policy makers in the near future.

Another interesting area is the study of the evolution of the social network over time. Although Twitter's search mechanism does not allow querying the exact moment of an edge creation, it is possible to reconstruct the social graph of a given time by applying various heuristics. Computationally this requires the development of algorithms to download an adequate and practically feasible sub-sample of Twitter's data, construct the dynamic social network and perform the graph analysis. With this we can travel back in time like rewinding videotape and study the social graph while it is expanding. This will help us to study how Twitter penetrated specific geographic regions, or online communities with no spatial ties. Finally, it can even help us to locate anomalies in the graph that are suggestive for spam or malware activity.



The first plot is Max Nanis '12: Community Detection on My His Personal Twitter Account, the middle one a Network Prior to Bot Activity and the last one the Same Network After Bot Activity
From: Published annually since 1943, The Silo is a student-run and produced journal of arts and letters at Bennington College in Bennington, Vermont: November 2011 Prose, Visual, Volume 68 Issue 1, Max Nanis, Prose, Silo, Visual, Volume 68, Volume 68: Issue 1
http://silo.bennington.edu/the-topologyof-human-networks-online/

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

Ecological Theory of the Digital World

The success of Web 2.0 has changed the way humans interact on a worldwide scale through online social networks, which nowadays connect more than one quarter of the world's population. The success of these networks is based on their capacity to engage users whose time is a limited resource. This necessarily leads to competitive interactions between the different digital services. The digital world thus forms a complex ecosystem of interacting networks. Analogously to the case of population dynamics, the question: “Why do we observe a moderate number of coexisting digital services?" arises and remains unanswered. Our results explain how multiple networks can coexist and why we only observe a moderate number of coexisting services in the digital world.

A rich-gets-richer mechanism impedes the coexistence of species in competition for the same resource according to the exclusion principle. We find that the dynamics of networks [1], however, induces diminishing returns which damp this effect, hence allowing a stable coexistence of several networks [2]. The stochasticity of the underlying processes and the form of the basin of attraction limit the observed diversity in the digital ecosystem in accordance with empirical observations. The general level of observed diversity is controlled by the influence of mass media.

Just as a monopoly in economy is a threat to free markets, the lack of digital diversity poses a threat to the freedom of information.

Our work provides the foundation of an ecological theory of the digital world. We suggest additional studies be conducted to include extensions such as different network intrinsic fitnesses. Enriched with empirical data, our theory can be extended to a description of the worldwide ecology of OSNs by incorporating different underlying societies that represent different countries.

References:

1] Kaj-Kolja Kleineberg and Marian Boguna, Evolution of the digital society reveals balance between viral and mass media influence. Phys. Rev. X, 4:031046, Sep 2014.

[2] Kaj-Kolja Kleineberg and Marian Boguna. Ecology 2.0: Coexistence and domination among interacting networks. ArXiv:1410.8865, 2014.

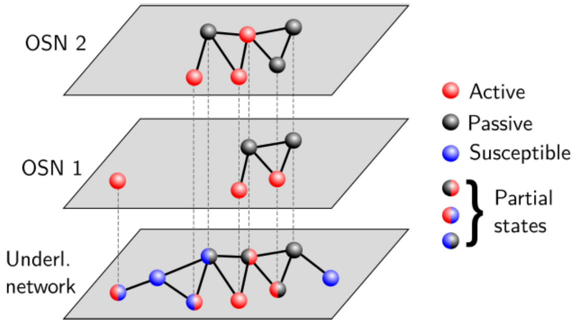
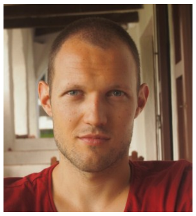


Figure 1: Multiplex layout for description of the digital ecology

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



Identification of Key Locations based on Interactions in Online Social Networks

Online Social Networking is one of the main means of communication in our era. A person is able to actively participate and interact with such platforms by sharing or reading information that other users share. Additionally, with the use of smart mobile devices and ubiquitous Internet connectivity, a user is able to publish information, nomatter the place that she is located. Thus, the information that a user shares in these platforms is enormous, providing personal and collective insights and predictions. The research community is highly interested in how all this information that is generated through online social networking interactions, can be used in order to extract useful knowledge that will help us improve our daily activities. Researchers analyze this information for several applications such as events detection, recommendation and early warning systems, personalized news extraction.

In this research we aim in identifying a user's Home, Work and Leisure locations, at post-code level, with the use of Online Social Networks data.

A person's base geographical location is a characteristic that can provide us with insights about both the individual and the location. These locations could be her home, work or leisure places and knowledge about them can provide us with interesting insights and potentials for recommendations. For example, knowing the characteristics of an area, a system can recommend it to someone who is looking for an apartment to rent and has specific interests and background.

Moreover, knowledge about leisure areas where habitants of another area spend their time could also be of interest for persons who live in an area with similar characteristics. A main problem in these scenarios is that only a small number of users provide accurate information about their base locations. In this research, we name these areas as *key locations* and we aim in identifying them with the use of Online Social Networking data.

When a user posts a text or an image on her profile, she does not share only the content but also a meta-data part, which in some cases is more important than the actual content. These meta-data include, among others, the geographical coordinates of the place where she was physically located when she posted the information, along with a time-stamp, which denotes the exact generation time. All these temporal geo-tagged information sketches users' mobility trajectories and pinpoints their visited locations. By analyzing these generated trajectories we are able to infer user's home, work and leisure areas. With this knowledge we are able to profile the different areas based on their habitants characteristics. In our work we aim in taking advantage of this geo-tagged data and propose an effective methodology for identifying a user's key locations, namely her home, work and leisure locations at a post-code granularity.

For this research we use a user's geo-tagged Twitter activity traces to identify her key locations, namely her Home, Work and Leisure areas. Our method is based on two basic observations. First, users tend to spend a significant, but distinct, amount of their time, during an average day in their Home and Work locations. For example most of the weekday evening hours users tend to reside at home. Second, these two locations are much more likely to appear in the users geo-tagged activity during these specific timeframes, than locations that are not embedded to the user's routine. Based on these insights we propose a model, which is able to improve existing methodologies. Our results show that our method can accurately identify the user's key locations, with precision values close to 80%, in post-code granularity. This result is not only an improvement of 30% over the state of the art, but also offers key location identification in a much fine grained granularity over the 10Km and city level range identified in previous work.

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

Building a Smarter Social Network

Social Networks have succeeded in enhancing online interactivity among people; however the full benefits of social networks are not entirely exploited. Therefore, Decentralized Social Networks (DOSNs) are proposed with the potential to re-consider the well-being in providing services closer to users' needs and expectations. DOSNs give the authority back to the users by allowing them to orchestrate the main building blocks that control data access and application management.

We propose Smart Social Networks (SSNs) with the potential to engage more effectively and actively with users to bring the peak benefit of big data.

Moreover, users will be able to perform different reasoning tasks by applying machine learning algorithms (Figure 1). In particular, such reasoning tasks are designed to make social network applications smarter, easier and better. It's not just about making accessing services more convenient, we are working on building a smarter future for social networking at large, and big part of our vision is secure and privacy preserving environment. Therefore, we propose Smart Social Networks (SSNs) with the potential to engage more effectively and actively with users to bring the peak benefit of big data. SSNs enable modeling users behavioral patterns, thus provide better personalized and context-aware services.

For example, we have finished a joint research project with University of Insubria to develop a decentralized identity validation model. The core idea of our proposed model is to mine the correlations among user profile attributes inside every individual community, where the correlations are more pronounced. In particular, identity validation patterns are deduced from these correlations and are used as guidelines that users can use to evaluate the reliability of the new profiles they wish to interact with. Our model operates in a fully decentralized manner and experiments using real-world OSNs datasets show that our model is able to extract fine-grained community-aware correlations among profile attributes with average improvements up to 50% than the centralized approach by considering the whole user population at once (Figure 3).

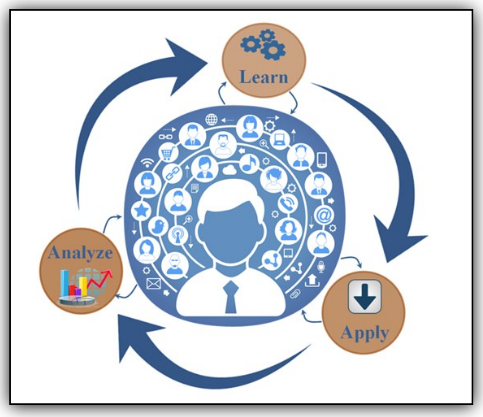


Figure 1: Integrating the users into the incremental learning process

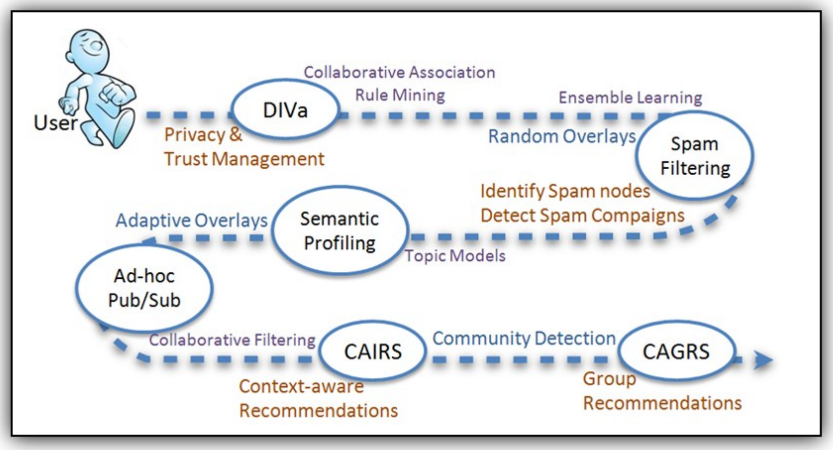


Figure 2: A roadmap example of smart services that can be offered in DOSNs

Furthermore, we target to incorporate deep learning approaches with topic modeling to build adaptive topic modeling techniques that work on large, real time social text streams. In particular, adaptive topic modeling automatically groups incoming social feeds based on semantic similarity. These groups or topics can then be used to provide scalable indexing, categorization, and retrieval techniques. Current approaches of topic modeling are computationally expensive especially when the data itself is dynamically growing at a high rate, which is the case of social networks like Twitter, and Facebook. Therefore, we plan to use a set of models created based on detected communities in social graph. Furthermore, we need to extend our algorithm to be an online inference algorithm for topic models, whereby the representation of the topics in a community is incrementally updated by new feeds added by users belonging to it.

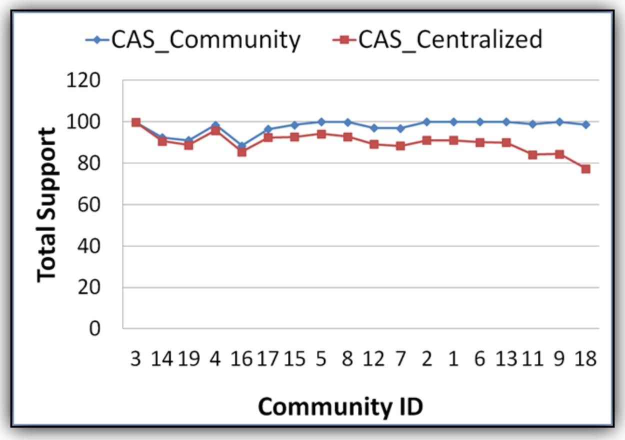


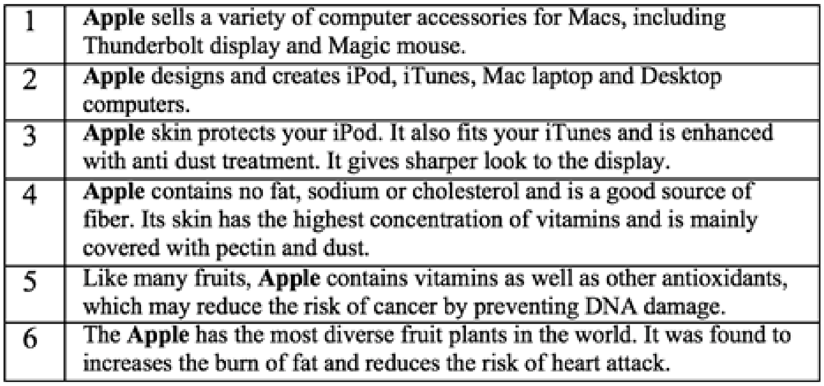
Figure 3: Comparing our model results (CAS_Community) with the centralized results (CAS_Centralized)

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

Distributed Semi-Supervised Multiple Disambiguation

Disambiguation is the task of determining the true entity behind an ambiguous mention in a document. Humans perform this task by manual assessment of unambiguous surrounding context words. However, manual disambiguation is not a solution when we are facing web scale documents. Current proposals use similarity based clustering for automatic disambiguation of a large number of documents. Graph-base community detection, recently proposed in our group, is one such solution that resolves scalability issues by applying a distributed processing approach. Each sentence in Table 1 contains the ambiguous word "Apple". Our solution first, transforms the documents into a graph, such that similar documents construct dense areas (communities) in the graph (Figure 1).

We proposed a solution for disambiguation of multiple ambiguous mentions, Multiple Disambiguation (MD) which is also, based on distributed community detection over graphs .



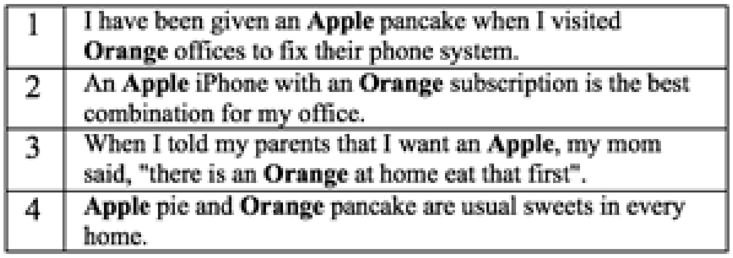
1	**Apple** sells a variety of computer accessories for Macs, including Thunderbolt display and Magic mouse.
2	**Apple** designs and creates iPod, iTunes, Mac laptop and Desktop computers.
3	**Apple** skin protects your iPod. It also fits your iTunes and is enhanced with anti dust treatment. It gives sharper look to the display.
4	**Apple** contains no fat, sodium or cholesterol and is a good source of fiber. Its skin has the highest concentration of vitamins and is mainly covered with pectin and dust.
5	Like many fruits, **Apple** contains vitamins as well as other antioxidants, which may reduce the risk of cancer by preventing DNA damage.
6	The **Apple** has the most diverse fruit plants in the world. It was found to increases the burn of fat and reduces the risk of heart attack.

Table 1: Sentences including ambiguous word "Apple"

Then, the algorithm determines communities through local communication between nodes of the graph. The algorithm starts with as many communities as the number of documents in the data set. Then, nodes engage in a local iterative process where everyone tries to merge those categories, if they happen to be in topologically similar locations. The location information is implicitly conveyed through local information propagation.

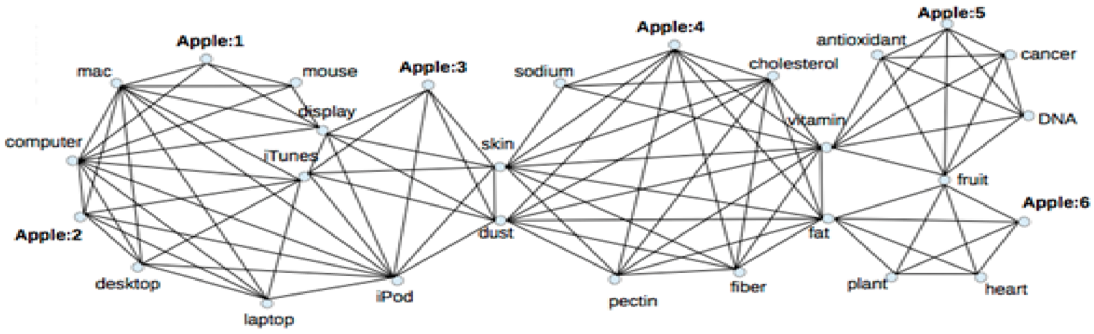
The previous model, also known as Single Disambiguation (SD), relies on a strong assumption requiring that every document contains a single ambiguous mention and a rich set of unambiguous context words. This assumption is valid if documents are carefully collected from a specific topic, however, real disambiguation problems, today, are engaged with documents collected from various sources like news feeds, reviews or twitter posts that often contain more than a single ambiguous mention (Table 2).

Applying SD to such data sets, a distinctive run for different ambiguous mentions is required, which on one hand, imposes a large number of extra processing steps for every single disambiguation, and on the other hand, detracts the quality of the results by removing available information.



1	I have been given an **Apple** pancake when I visited **Orange** offices to fix their phone system.
2	An **Apple** iPhone with an **Orange** subscription is the best combination for my office.
3	When I told my parents that I want an **Apple**, my mom said, "there is an **Orange** at home eat that first".
4	**Apple** pie and **Orange** pancake are usual sweets in every home.

Table 2: Ambiguous words included on various documents



We proposed a solution for disambiguation of multiple ambiguous mentions, Multiple Disambiguation (MD) which is also, based on distributed community detection over graphs. However, we changed the graph representation to account for MD (Figure 2). As we can see the graph in MD is more complex compared to SD. Our experiments show that this complexity causes wrong classification of ambiguous mentions referring to multiple entities from the same topic. For example, the ambiguous mention "Clinton", referring to either "the Hilary Clinton" or "the Bill Clinton" entities, often categorizes into the same cluster, since both entities belong to the same topic.

Therefore, a new parameter was developed to prevent such miss-categorizations by controlling the resolution of the clustering. In fact, this parameter is used to extract more confident results by having a large number of high quality clusters (high precision) rather than a small number of low quality clusters (high recall). First, we tune the number of clusters in a way that the clusters still have high precision and they are not too many so that we can manually detect the topics for each cluster in a scalable manner. Then, we merge those clusters that fall in the same topic by relying on preexisting knowledge of the ground truth or by manual investigation.

Experiments show that few samples increase the recall while maintaining high precision. As a prototype, applying the algorithm over a synthetic data-set, extracted from around 2K Wikipedia with 20k nodes and 1M edges in graph representation, resulted in 62% recall while maintaining precision on 75% with sampling only 4% over the whole data-set.

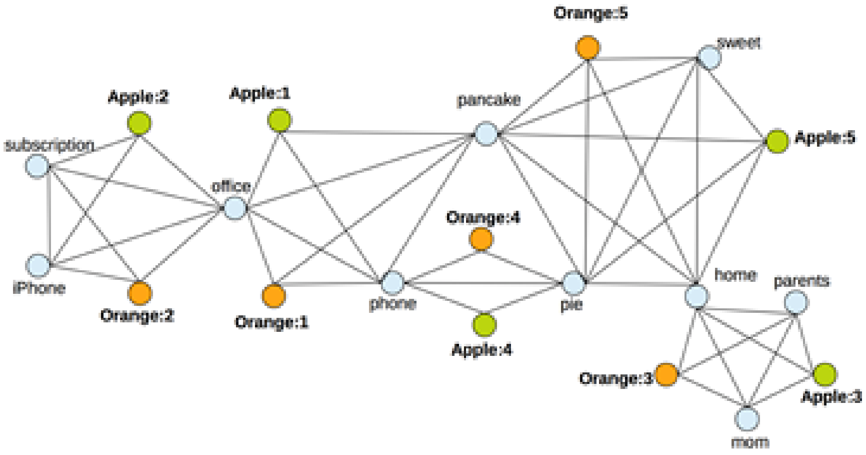


Figure 2: A graph representation to account for MD

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

Network Modeling and Overlay Design for DOSNs

An important aspect of designing a decentralized online social network is the specification of the overlay network, the topology of connections between peers that are used to manage the services provided by the platform in a decentralized manner.

It is natural to utilize the social network in the overlay design, as one's social connections in the platform will be the nodes with which one most often exchanges messages. At the same time, in order to be able to efficiently search for information and people on the platform, overlay connections must be maintained between machines belonging to users who are not friends, or even close to each other, in the social graph.

In the context of utilizing the social network model for overlay design, a need to model the social network layer and to study its various properties becomes apparent.

In an iSocial research project, a collaboration with Stefanos Antaris, Mikael Högqvist, George Pallis and Marios Dikaiakos, we consider these two competing objectives to design a balanced socially-aware overlay network. We build upon existing platforms designed to implement peer-to-peer (p2p) distributed hash tables (DHTs), and propose an augmentation that preserves guarantees for fast search and exchange of messages in existing systems, while leveraging features of the online social network application, to reduce these in many cases.

In the context of utilizing the social network model for overlay design, a need to model the social network layer and to study its various properties becomes apparent. This brings to focus Network Science, utilizing theories and tools from various disciplines including Computer Science, Statistics, Sociology, and, of course, Graph Theory. One of the goals of Network Science is to design and apply statistical models for networks, proposing a distribution assigning a probability to each possible graph or network. This allows one to analyze network data, and perform statistical inference, in order to conclude that certain properties should be expected under a model, as well as to produce simulated networks with similar properties as the original, or to make other inferences.

Exponential Random Graphs are some of the nicest statistical models for networks, in their simplicity and in that they inherit some very convenient mathematical and statistical properties from exponential families of distributions. In prior and ongoing projects with collaborators, we have proposed, studied and utilized Exponential Random Graph Models (ERGMs), for directed networks, an example of which would be Twitter; an ERGM based on the core decomposition of a graph, which gives a notion of importance of the nodes in the network; and an ERGM to model groups using hypergraphs, generalizations of graphs.

Despina Stasi
despina.stasi@cs.ucy.ac.cy
ER iSocial Fellow
University of Cyprus (UCY), Cyprus

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

Newsletter Content Editor:
Kalia Orphanou
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)

https://www.facebook.com/ISocialMarieCurieITN