Issue 1, May 2014

iSocial: Decentralized Online Social Networks

Inside this Issue

Trending Topics on Decentralized Online

Decentralized Social Networking Platforms: Current Status and Trends

Data ownership and privacy issues in Online Social Networks are two factors that have motivated the research community to look into distributed systems as a possible solution. The problem is that distributed systems are complex and in the case of social networks, sensitive information can be exposed. In this article, we present an overview on the current status of the ongoing challenge to decentralize Online Social Networks.

read more

RankSlicing: A Decentralized Protocol for Supernode Selection

In peer-to-peer applications it is common to assign greater responsibilities to peers having higher computational power. The presented algorithm offers a practical solution for selecting a set of K peers for the super-peer role.

read more

Hive, a Distributed Caching Network

In 2012, video traffic on the Internet, including both Video on Demand (VoD) and live streams, was more than 57% of all IP traffic and is expected to grow to 69% by 2017. Hive is a Distributed Caching Network that has been shown to consistently save more than 90% of the traffic to the source and on bottlenecks in the network.

read more

User Privacy in Online Social Networks

The impressive adoption of social networking services and their ease of communication raises many concerns regarding users' privacy. Users often tend to share personal and private information without hesitation as they are unaware of the implications of their actions. Thus, it is imperative to raise user awareness about possible threats and the importance of privacy.

read more

Risk Assessment in Decentralized Social Networks Based on Anomalous Behavior Detection

Risk analysis and trust management in decentralized social networks are essential and important elements for successful social networking experiences. This project aims at investigating mechanisms for the detection of risky users. This is achieved under the assumption that the more the user behavior diverges from "normal behavior" the more he has to be considered risky.

read more

Validating OSN Users' Identities Using Community Feedback on their Profile Values

Motivated from some sociology findings on identity formation, we have explored the possibility of evaluating OSN profiles' homogeneity based on community feedback. The objective is to use this measured homogeneity as an indication on the truthfulness of the claimed online identity. Our initial experiments proved promising.

read more

Situation Aware Social Overlay

Online social media have been established as the main platform for information diffusion. The enormous data that can be found out there provide the public with the ability to access information that lie in their interests. However, not all publishers or data are of the same interest at all points in time. On one hand, if a user is not connected with the source of the information he will not be able to access it. On the other hand, if a user is connected with many resources he will face the problem of information overload. With this research we aim in connecting people with the information that are interested in, taking into account the current situation, removing the necessitation of directly connecting users between each other .

read more

Mixing Models for Better Learning

Ensemble Learning – also known as mixture of experts, uses multiple classifiers to obtain a better prediction performance compared to single classifier systems. In such systems, a set of learning nodes are exchanging intermediate classifiers and then train a global classifier using all meta-level training subsets. Our research targets building Ensemble Learning algorithms for Decentralized Online Social Networks.

read more

The Evolution of the Structure of Online Social Networks

The rapid growth of online social networks is reshaping the social landscape changing the way humans interact on a world-wide scale. Our study reveals and quantifies very precisely the principal mechanisms underlying the structural evolution of online social networks.

read more

Large Online Social Networks

Online Social Networks occupies the major portion of Internet traffic. Due to the emergence of various handheld devices, the numbers of users on different OSNs are exponentially increasing. This OSN adoption trend creates various scalability and management challenges for the OSN providers, like service availability, fault resilience, and data security. However, these challenges can be refuted at the cost of expensive resources. Therefore, there is a strong need for resource efficient solutions that can handle the increasing trend of OSNs .

read more

Exploitation of Trending Topics on Twitter

Twitter is a fast growing and popular Online Social Network based on a very simple data model. This simplicity attracts, apart of hundreds of million of users, spammers, phishing attempts and malware. The same principles and features that made twitter a huge success are also exploited for promoting malevolent content. This work is towards uncovering these exploitation patterns and build defense mechanisms.

read more

Apple Skin Protects your iPod!

Cross-Document Reference Resolution is often defined as an NP complete similarity based categorization problem. Numerous extensions based on vector space modeling (VSM) structure have been proposed, which mainly disregard its inherent inefficiencies caused by the number and the size of similarity comparisons it imposes to the clustering algorithm. We propose a scalable solution based on an innovative graph based modeling structure and a content agnostic, diffusion based, node-centric clustering algorithm.

read more

Decentralized Social Networking Platforms: Current Status and Trends

Social networks have changed dramatically the way in which people interact. The information we share has become so important for companies and individuals that multibillionaire businesses have flourished around these ecosystems. Online services such as social networks, cloud storage, micro-blogging and video sharing provide rich ways to interact with billions of users around the globe. Such services do not come completely for free. Recent developments have shown that our data is being monetized and monitored as we speak. Additionally, censorship and the lack of data transparency are constant barriers in these services. We believe that as a society, we need a more open digital world in order to express ourselves without losing our right to privacy. Research efforts, open source projects and active communities have been trying to innovate towards this idea. We found that existing alternatives aim to alleviate privacy and data ownership concerns using a variety of protocols and architectures. We present a classification of these projects in **Figure 1.** The color code denotes the protocols the project uses and the position of the label represents the maturity in terms of release stage (alpha, beta or stable).

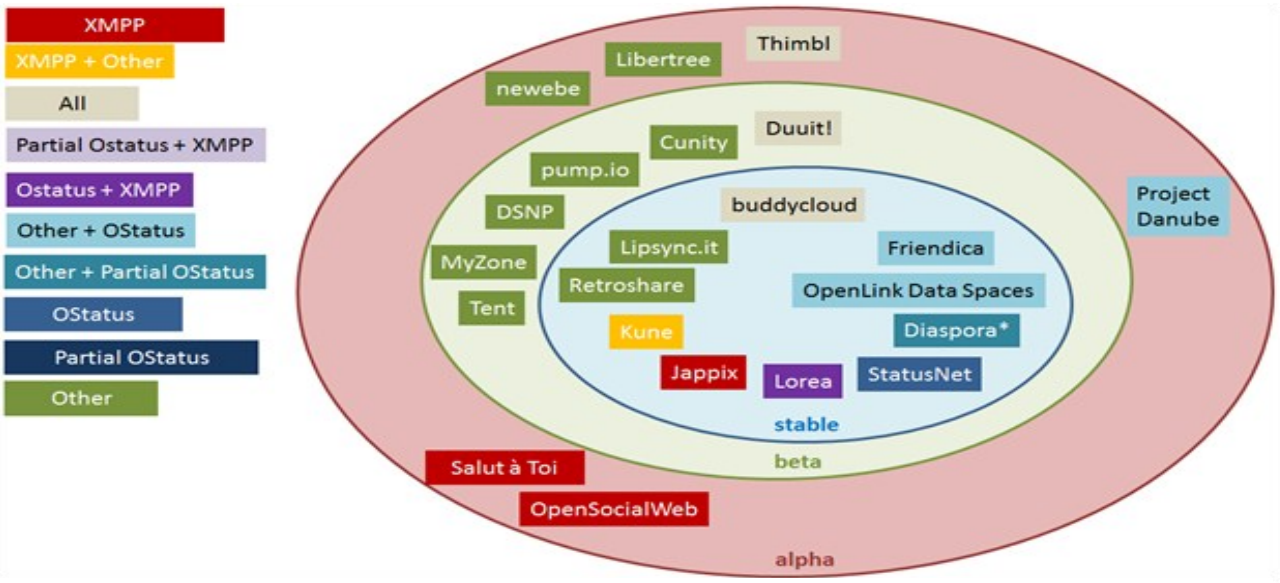


Figure 1: Taxonomy of existing projects that aim to decentralize Online Social Networks. The figure shows the protocols that they use and the release stage of the project.

After looking at the current work from communities around the globe, we classify these projects into two approaches: Peer-to-peer and Federated Systems. **Table 1** summarizes their advantages and disadvantages. Current technological breakthroughs, in software and hardware, could overcome these limitations. For instance, a low-cost device, such as Raspberry Pi, could run a light-weight software stack able to provide similar online services to a small number of users.

Type of Distributed Social Network	Advantages	Disadvantages
Peer-to-peer: PeerSON Safebook Retroshare	**Privacy**: Each peer shares whatever information the user decides, the privacy is in the hands of the user. **Data ownership**: Each peer can withdraw its data and keep it offline; there are no restrictions as to what an owner can do with it.	**Availability**: Peers can go offline making its data unavailable. **Performance**: The performance of a peer is comparatively lower than that of a dedicated server in terms of CPU, RAM and bandwidth.
Federated: Diaspora StatusNet	**Availability**: A dedicated server is more likely to be online 24 hours than a peer that comes and goes off the network. **Performance**: A dedicated server usually has higher CPU, RAM and bandwidth than a peer. Consider that someone needs to pay for these expenses.	**Privacy:** The data is stored in a server controlled by a third-party. Thus the administrator could access private conversations if the architecture of the service allows it. **Data ownership**: Since the data is stored in a third-party service, the user no longer controls it. That is, it cannot be deleted, unless the administrator explicitly does it.

Table 1: The two approaches that aim to decentralize Online Social Networks. The table shows some examples, their advantages and disadvantages.

A group of devices may comprise a network able to exchange data following strict user defined privacy rules. Further study on the feasibility of this idea requires more literature review and the use of simulators to test the capacities of these devices. However, their hardware specifications allow the creation of a personal digital space for each interested user (that is, a social network, a personal blog, an email server and a cloud storage service). Using such devices, but not limited to, we could overcome the limitations of peer-to-peer systems achieving higher availability. The data stored in the devices can be protected using encryption mechanisms enhancing the notion of privacy, thus overcoming the limitations of Federated Systems where administrators have unrestricted access to the data and any transactions. Additionally, the data is stored locally. This implies that the user has absolute control over what, when and how to share information. At a given moment, the same information can be put offline as the device is in the possession of the user, in contrast to cloud service providers where sites are located miles away from the data owners.

The next step is to explore the state-of-the-art technologies in order to build the envisioned infrastructure. After collecting sufficient knowledge, we will proceed to develop prototypes that will be expanded into full-fledged platforms for distributed social applications. Once we have a complete platform we will perform the corresponding tests and compare the results with other projects.

Andres Ledesma
aledesma@cs.ucy.ac.cy
ESR iSocial Fellow
University of Cyprus (UCY), Cyprus

RankSlicing: A Decentralized Protocol for Supernode Selection

Video streaming has become an important service in today's Internet. It is a very common practice for users within social networks to share to multimedia content. The infrastructures for video propagation require, however, a more careful design with respect to other kinds of multimedia, as it is much more resource demanding.

Peerialism AB is one of the leader companies in the field of distributed video streaming, leveraging the Peer to Peer technology for providing a cost and bandwidth effective service.

Within this context, we are conducting a research project on the design of a decentralized k-Leader Election algorithm named RankSlice, which aims to be an important building block for the technology implemented by Peerialism. The k-Leader Election problem consists in selecting a subset of k super-nodes to be assigned to a certain special role, basing on their better performances.

In the peer-to-peer video streaming application the goal is the identification of the better nodes in a subset (e.g. a LAN) so that they can act as content providers for the other nodes

The algorithm is very general by design, and allows a distributed application to specify the relevant system characteristics for being a leader. In the peer-to-peer video streaming application the goal is the identification of the better nodes in a subset (e.g. a LAN) so that they can act as content providers for the other nodes.

Besides content distribution, which is part of the technical infrastructure of a Distributed Online Social Network, multiple use cases for the k-Leader Election problems can be identified on the social layer, like for instance, the automatic identification of a subset of people, within a users' group, which may be elected as moderators for a discussion on a bulletin board.

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

Hive, a Distributed Caching Network

Video as a communication medium is a critical part of online social networks such as Youtube, Facebook, Vine and Twitch.tv. In 2012, video traffic on the Internet, including both Video on Demand (VoD) and live streams, was more than 57% of all IP traffic and is expected to grow to 69% by 2017. This creates large amounts of traffic going through the core of the network. Fortunately, each video is duplicated many times which makes it easy to cache video closer to the consumer by using distributed caching networks such as CDNs.

Hive is a software-only solution and is able to reach quality of user experience which is on-par with or better than CDNs, while minimizing redundant traffic in the network

Peerialism has developed a novel distributed caching network called Hive focused on live video within enterprise networks where CDNs have proven difficult and expensive to deploy and manage. In contrast to CDNs, Hive is a software-only solution and is able to reach quality of user experience which is on-par with or better than CDNs, while minimizing redundant traffic in the network. Distributing video efficiently is a core application within distributed online social networks, and used for example in teaching efforts or live one-to-many communication such as town hall meetings.

In work presented at **SIGCOMM 2013**, we showed a distributed caching solution which addresses the problem of efficient delivery of HTTP live streams in large private networks (**Figure 1**). With our system, we have conducted tests on a number of pilot deployments. The largest of them, with 3000 concurrent viewers, consistently showed that our system saves more than 90% of traffic towards the source of the stream while providing the same quality of user experience of a CDN. Another result is that our solution was able to reduce the load on the bottlenecks in the network by an average of 91.6%.

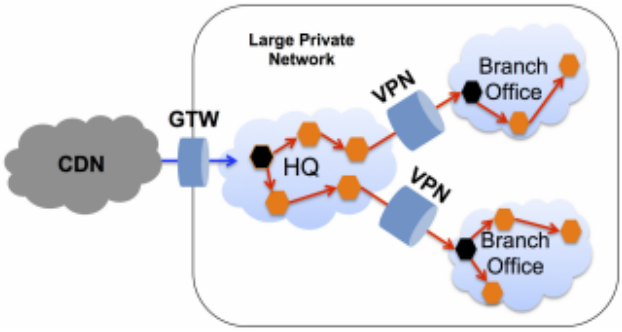


Figure 1: HTTP live delivery in a private network with our solution

Source: Roberto Roverso, Sameh El-Ansary, Mikael Högqvist: On HTTP live streaming in large enterprises. SIGCOMM 2013: 489-490

Mikael Högqvist
mikael.hogqvist@peerialism.com
ER iSocial Fellow
Peerialism AB, Sweden

User Privacy in Online Social Networks

In the last few years we are witnessing the rapid emergence of Online Social Networks (OSN), which in turn, have radically changed the structure and the utility of the Internet. The great success of social networking applications is reflected by their enormously increasing user-base. For example, Facebook has recently surpassed 1.2 billion registered users.

The success of OSNs mainly stems from their nature and inherent social character. In fact, social networking applications allow the users to create their own digital profiles and get connected with real-life friends and acquaintances. Thus, the users are able to communicate with their friends and even with other users under common interest groups.

Apart from the ability to communicate, the users usually feel the need to contribute to the community by providing reviews, suggestions and feedback. Moreover, the users can create and post audio-visual/textual content (such as photos, comments, tweets, videos) which is usually denoted as "Freedom of Speech". Recently, "Freedom of Speech" has been limited in countries such as China and Turkey by censoring websites and access blocking social applications such as Twitter, YouTube and Facebook.

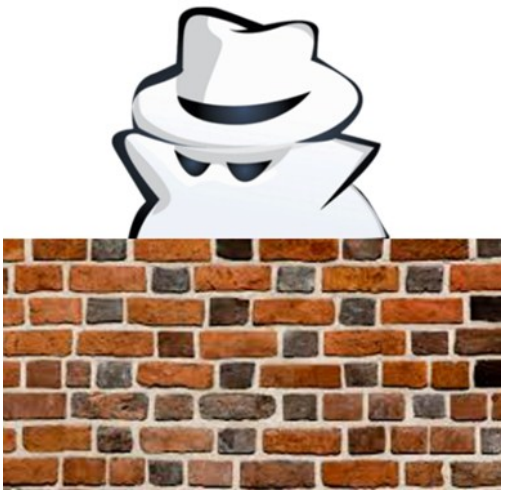
If you feel like someone is watching you, you're right. If you're worried about this, you have plenty of company. If you're not doing anything about this anxiety, you're just like almost everyone else (Bob Sullivan)

However, the impressive adoption of social networking services and their ease of communication raises many concerns about users' privacy. It seems that the average user is not concerned about his/her privacy and often tends to share personal and private information without hesitation. This information can be related to user location, financial status, political and religious views, sexual orientation and so on, which can potentially lead to cyber-bullying, social harassment and crime. Nevertheless, the above issues have not yet been discussed in a large scale.

Moreover, in the case of photo uploading, each user "*tagged*" in a photo can literally affect its visibility by allowing his/her friends to access it. In this case, both the photo up-loader and the "*tagged*" users are unaware of who is potentially viewing their photo.

Additionally, most of the users are unaware of the implications of their actions and the potential harm posed by publicly sharing data. It is a generally common practice for a number of companies to regularly look up the online profiles of their job applicants as a hidden part of the hiring process.

This practice is important for these companies as the online activity of the applicants can reveal aspects of their personality.



There are also many reported cases of users having to face dire consequences in their personal and professional life due to information they had recklessly published.

Despite the carelessness of the users' social behavior, another important factor affecting user privacy is the current design approaches followed by social networking providers. Usually the mechanisms employed for controlling data accessibility is either too simplistic (e.g. access to none, all friends, everyone) or too complicated for users to comprehend and comply with. The non-decentralized design of the current social networks is also affecting users' privacy as their personal information is permanently stored within the providers' servers. At the same time, the providers can track user interaction and can collect information about their online activities. This information can easily be used by the SN providers or by third parties for revenue purposes, such as advertisement.

If you're using Flickr or Delicious or YouTube or belong to Facebook or LinkedIn or another of the popular social networks, you've given up complete control of your personal information. You don't, so to speak, "own" it anymore. Surprised? (Debbie Weil)

It becomes apparent that there is an imperative need to preserve the privacy of the users while they are using social platforms. The security and the privacy issues in social networks need to gain the attention of the traditional media in order to initiate a discussion regarding the risks and raise user awareness about the importance of privacy. Furthermore, we need to design and build new simple mechanisms for fine-grained information sharing. It is very important for each user to be able to control his personal information and data.

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

Risk Assessment in Decentralized Social Networks Based on Anomalous Behavior Detection

Decentralized Social Networks allow users to create a public or private profile, encourage sharing information and interest with other users and communicate with each other. Therefore, users interact with each other both socially and professionally in the virtual environment. However, both in traditional social networks and in decentralized ones, users are used to establish new relationships with unknown people with the result of exposure of a huge amount of personal data. Unfortunately, very often users are not aware of this exposure as well as the serious consequences this might have. Some of the information posted on these sites can lead to security risks such as, identity theft and cyber stalking. Therefore, to get a safer environment, risk analysis and trust management in decentralized social networks are an essential and important element for successful social networking experiences.

Both in traditional social networks and in decentralized ones, users are used to establish new relationships with unknown people with the result of exposure of a huge amount of personal data

At this purpose, as first stage of research activities in I-Social project, we made a comprehensive and detailed review of works related to trust and risk concepts in the context of social networks. This pointed out that mechanisms for risk/trust estimation have to be improved and adapted in order to be adopted in decentralized social networks. Based on this detailed review, we can conclude that available trust evaluation models can be organized according to three categories:

(1) Network-based models, where the network structures affect the level of trust of social networks users. High density in a network (higher interconnectedness between members) can yield a high level of trust. For instance, Increases in both the in-degree and the out-degree in turn increase the level of trust a member can have in another member. Therefore, receiving information from members with a higher in-degree increases the level of trust.

(2) Interaction-based models where social trust is computed taking into account patterns of interactions in the network. They consider interactions in the community for computing trust. Such interactions, for instance, are, the number of likes, comments, post and share items the user has, the average popularity of items that the user liked, posted and shared, the average time difference between posts, comments and share items and the percentage of received information user propagates to others.

(3) Finally, hybrid trust models where both interactions and social network structure are used to compute social trust.

Mechanisms for risk/ trust estimation have to be improved and adapted in order to be adopted in decentralized social networks

As such, the more the user behaviour diverges from "normal behaviour", the more he has to be considered risky. Thus, our goal is to detect anomaly in users by considering user's interaction with all other users in the network. This implies, at first, defining what the normal behaviour is, then detecting anomalies. We achieve these goals using clustering techniques that allow us to analyse and shape normal behaviour, and to identify anomaly in users with respect to emerged behaviours based on outlier detection .

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

Validating OSN Users' Identities Using Community Feedback on their Profile Values

Motivated from sociology with regard to its results on the convergence of developing identities to coherent wholes aligned with some socially accepted models, we have explored the possibility of evaluating OSN profiles' homogeneity based on community feedback. The goal is to deduce from this evaluated homogeneity an identity trustworthiness level to assign to a target profile. The objective of the work is to help OSN users filter the potential profiles they are to connect with based on their estimated trustworthiness, and to give them a means to determine the reliability of their virtual connections.

Our research work started by studying the possibility of capturing the truthfulness of a profile based on a human-judged coherence between some of its sub-parts. Each of these sub-parts corresponds to a group of profile attributes within which community apprehended some level of intra-correlation. For example, our experimental works revealed that, within our studied community, there is an intra-correlation between the educational major of the person and the sports that they tend to prefer. This intra-correlation derived from direct community feedback on several profiles in a training set. Our results show that this intra-correlation can be leveraged on to partially judge the coherence within the values of a given profile, from this community, with considerable reliability. In our initial experiments, our community-feedback based method scored an accuracy of 83% in rating the reliability of profiles.

Within our studied community, there is an intra-correlation between the educational major of the person and the sports that they tend to prefer

We believe that identity validation within the online social networks' arena is a critical and an important research field which needs more attention and deeper focus. Different types of identity attacks and breaches are in continuous development and honest end-users are always bound to accept the associated risks. To the best of our knowledge, our work is a first in addressing this identity validation issue from a community-sourcing perspective with the aim of helping users define their level of trust in the claimed identity within an OSN profile. We plan to extend our work over multiple dimensions and we look forward having it deployed as an eventual application within some popular OSN.

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

Situation-Aware Social Overlay

Online social media provide the user with the ability to interact, express and share his/her opinions, feelings and interests with the outside world; his/her family, friends, colleagues, neighbors. These interactions are not just users' statements but they hide much more valuable information behind them. Users' can now act as human sensors through their published messages in cases such as events detection, business environment analysis, real-world models (transportation, online behavior analysis etc.), social science investigation (psychology, trends etc.).

A user interacts with a variety of online services and shares interesting information. This information can be related with his opinions, interests, feelings, events happening close to him etc. Shared information can also lie in the area of interests of another user who is completely disconnected with the publisher

Another key aspect of our era is the enormous popularity of smart devices. The majority of people hold at least one smart device, similar to a smart phone or a tablet. These devices are equipped, or have the ability to connect, with a variety of sensors and give us the ability to collect data and use them for investigation in different fields.

Social networks are designed to provide social functionalities to the users such as connecting and/or interacting with other users with whom they share a level of relationship (real-life relationship, common interests, etc.). With this research, we aim at not only connecting users and devices based on their social-real connection but also based on their situation; we aim at building a situation-aware social overlay. We believe that a user is not only interested in what his/her friends share but also in what is happening in the outside world and is related to him/her, like events happening near his/her location.

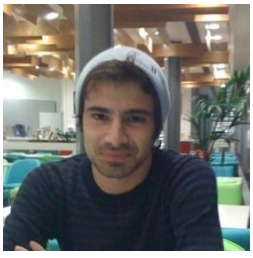
We aim at designing an overlay that gives us the opportunity to extract knowledge from human sensors (like the interaction in online social networks, articles in online social media) but also from device sensors (like Bluetooth sensors, smart device sensors, TVs etc.). Our goal is the design and development of a situation-aware social overlay that intelligently senses and provides personalized recommendations to the user, based on his/her interests.

We take into consideration that a user interacts with a variety of online services and shares interesting information. This information can be related with his opinions, interests, feelings, events happening close to him etc. As we can understand, shared information can also lie in the area of interests of another user who is completely disconnected with the publisher.

In the current state of Online Social Networks, this information remains hidden from the group of users who are not "socially" connected with the publisher. To better explain this problem we will use a scenario: Bob and Alice are Twitter users who live in the same neighborhood. However, despite the fact that they know the existence of each other, they do not follow each other. Bob went to the neighborhood's supermarket this morning and noticed that everything is half priced. He then tweeted the following message: "Feeling happy! I bought the supplies for the week in half price… My neighborhood's supermarket has everything half price for today!" Alice's regular plan is to visit the supermarket the day after. If she knew that the supermarket has offers, she would reschedule her visit in order to save money. However, the information that Bob shared reached his community (real-world friends, users with same interests) but Alice who was interested in this information didn't know its existence.

Our situation-aware social overlay aims at extracting knowledge from social networks, social media and devices in order to provide the user with situation-aware suggestions according to his/her daily life and interests considering the requirement of fast data: knowledge should be delivered to the user before it expires. In the case of our scenario, if Alice gets the notification the day after, then the information stops to be valuable and does not fall within Alice's interests. The identification of user's interests from online social networks-media profiles and interactions is a challenging area. Another challenge that our research aims to address is the knowledge extraction from online services and the identification of the relation between a user and the extracted knowledge.

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



Mixing Models for Better Learning

When we have an important issue that is related to our financial, medical, or social aspects of our lives, we seek asking for different opinions before making our final decision. These different opinions open our mind to consider the issue from different angles and perspectives. In doing so, we assign weights to the individual opinions and combine them in order to reach the decision that is the most informed one. The process of consulting "several experts " before making a final decision has attracted a lot of research in computational intelligence community, where opinions in computational domain represent the outcome or the models of individual machine learning algorithm.

Our research objective is to enable these users' devices, also known as nodes in social graph, to be linked to different sources of information and/or learning models from the whole devices / nodes exist in social network

Our iSocial project main objective is to build a framework that provides users with a set of tools and protocols to build their own online social network. This way, users will not be obliged to upload their own data such as pictures and videos to any remote servers as the case now with Facebook and Google+, for example. Instead, users will contribute with their devices such that a network of trusted servers can be built using these devices. Our research objective is to enable these users' devices, also known as nodes in social graph, to be linked to different sources of information and/or learning models from the whole devices/nodes exist in social network .

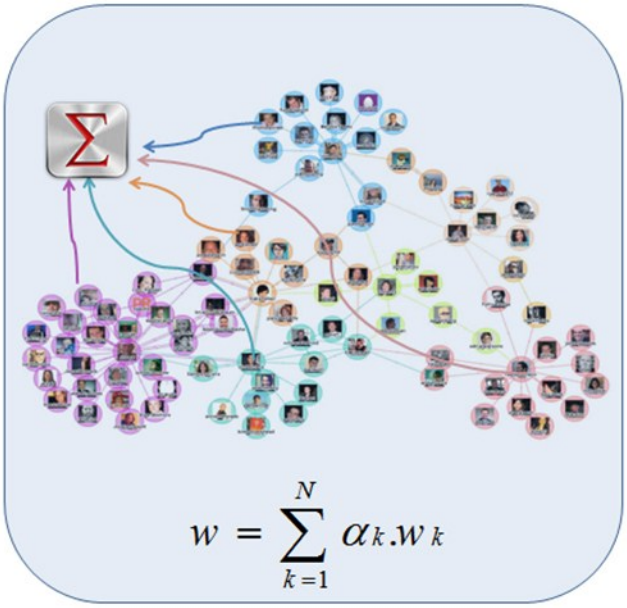


Figure 1: Merging models on top of social graph.

Our research work started by creating 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. While, in the second part of our research we study how to assign different weights to this set of learning models before combining them to get the final decision or learning model (**Figure 1).** We have developed our algorithm that can adeptly change the weights depending on the state of each individual learning model, for example the weight is increased or decreased based on size of information included in performing the individual learning process, and the time when this information joined the network.

Currently, we are testing our algorithms with different topologies of social network graphs and different data. Our objectives is to provide the users with adaptive set of functions such as link predication which means recommending some people to a user such that they can be friends in social network. Moreover, we also provide spam filtering to help users to detect spam users and messages and delete them from the network.

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

The Evolution of the Structure of Online Social Networks

The rapid growth of online social networks is reshaping the social landscape changing the way humans interact on a world-wide scale. Our study reveals and quantifies very precisely the principal mechanisms underlying the structural evolution of online social networks which are of huge interest for a broad spectrum of applications from viral marketing to optimizing personal success.

The vision of a decentralized Ubiquitous Social Networking Layer in the absence of central management

Our conclusions –virality being four to five times stronger than mass media influence, the coupling to the preexisting underlying social structure, and a higher viral tendency towards weaker ties– impact social, computer, and network sciences. Our work constitutes the foundation for the development of an ecological theory of the digital world comprising an entire ecosystem of online social networks competing for users' activity.

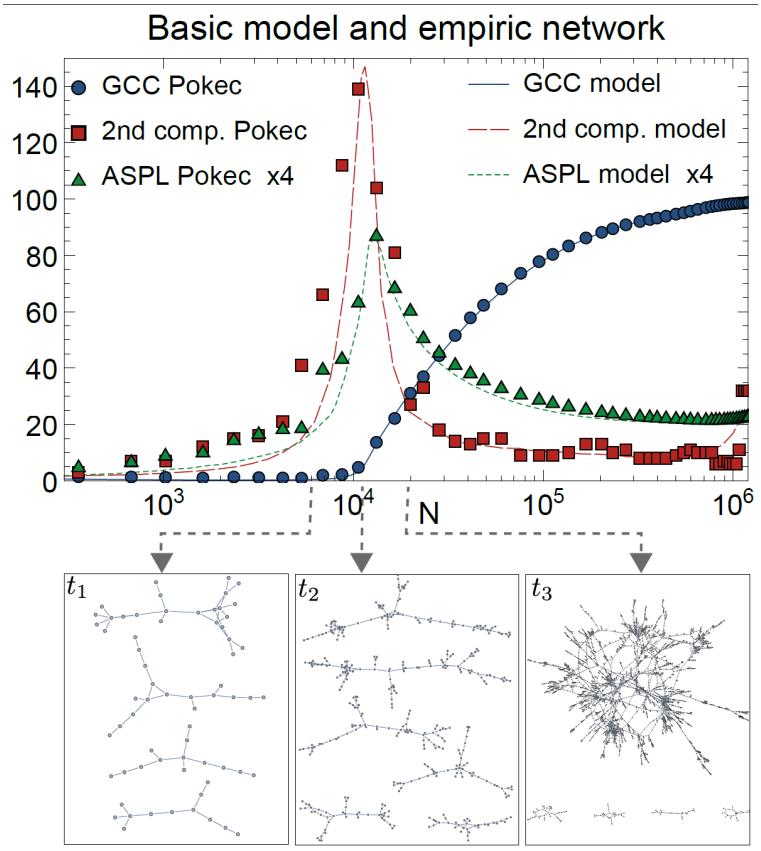


Figure 1: Topological evolution of the empiric network. Top: The inset shows the evolution of the network size from 1999 to 2012. The main plot shows the relative size of the GCC (blue circles), the size of the second largest component (red squares), and the average shortest path length (green triangles, multiplied by four for better readability). Bottom: The largest components of the network are visualized at three different times, before the critical point, t_1, at the critical point, t_2, and after it, t_3.

Source: Kaj-Kolja Kleineberg and Marian Boguna. Evolution of the digital society reveals balance between viral and mass media influence, May 2014, http://arxiv.org/abs/1403.1437

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

Large Scale Online Social Networks

Online Social Networks (OSNs) have been gaining magnificent growth and popularity in the last decade, as they have been attracting billions of users from all over the world and have been generating major portion of Internet traffic. Many well-liked companies, e.g. Facebook, Youtube, Twitter, Google, are investing resources to handle the huge workload that is generated by such networks. As a single machine cannot handle the amount of data produced by OSN, such networks require multiple machines to handle the workload. However, adding number of machines adds the complexity in the system. For example, data distribution in a data center environment requires an efficient technique to split OSN data and distribute it across the set of machines.

Any system that is designed to provide functionalities for online social networks should make sure that the user data is processed and stored in a secured manner

Handling the workload generated by OSN and storing it in a set of multiple machines is not trivial and requires special attention in order to utilize data centers efficiently. Moreover, we should understand that whether we use multiple machines or a single machine, the client or external user only cares about his data, which is related to his profile and all of data that is generated by his friends. Hence, the system should look like a single machine to an external client and the system should be able to handle failures of machines and other distributed system related issues without interfering a client. Similarly, systems handling social networks face additional problems that are related to the nature of social networks, like privacy and data security. Therefore, any system that is designed to provide functionalities for online social networks should make sure that the user data is processed and stored in a secured manner.

We can divide the systems that process OSN into two categories that are storage systems and analytic systems. Both these types of systems require data centers for their deployment. One of the naïve ways for data distribution in such an environment is to use a random distribution technique, where you randomly assign each of the users to any of the machines in the system. However, in case of OSN, we want to place users close to their connected friends and using random placement of users will break social ties among users. Hence, there is a need for intelligent data allocation strategy that can place the OSN user across data centers without breaking the social ties, i.e, users are placed on the same machine with their friends. Similarly, there are various issues that are related with the nature of OSN. For example, in OSN there is a huge number of users with very few connections and there are some users with very large number of connections. Using a random assignment for this type of data will place each user with the same importance, and the famous users will end up generating more workload for some machines. Hence, there is also a need for an efficient scheme that takes care of skewness of the data and utilized the hardware efficiently.

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

Exploitation of Trending Topics on Twitter

It has bee called the SMS of the Internet. It has played pivotal role in uprisings that led to significant social reforms and has outraged oppressive regimes. It is one of the most common channels of communication between celebrities and their fanbases. It is a major social, educational and news-feed medium used by more than half a billion people and it is only eight years of age. With posts that travel faster than an earthquake twitter is a revolutionary online social network.

The messages cannot exceed 140 characters in length and everyone can follow anyone with no restrictions. While you were reading this text, around one hundred thousand tweets were exchanged. A fair portion of them was about an obnoxious Canadian teenager, but another fair amount was spreading news and updates about everything that moves on Earth, regardless of notability.

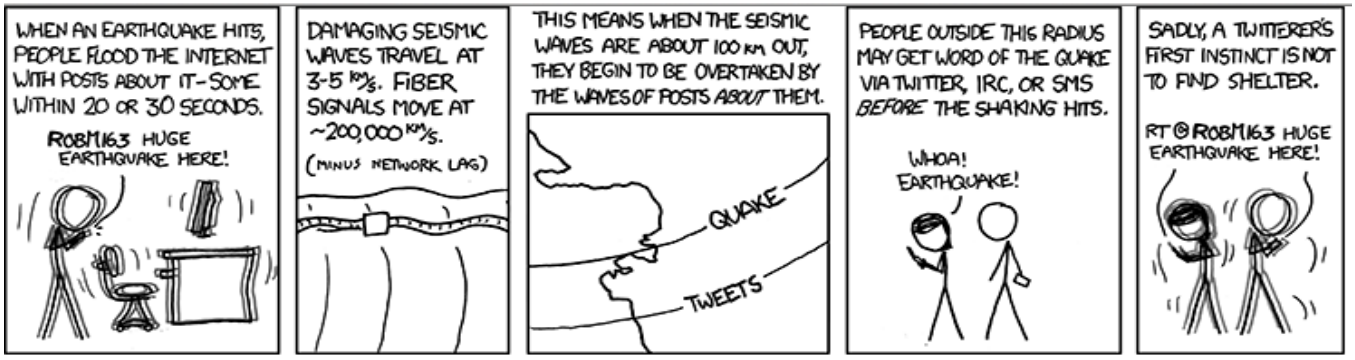
Is it just the ability to deliver short texts that make Twitter an addictive social network? Probably not.

Twitter provides us with a wonderful platform to discuss/confront societal problems. We trend Justin Bieber instead (Lauren Leto)

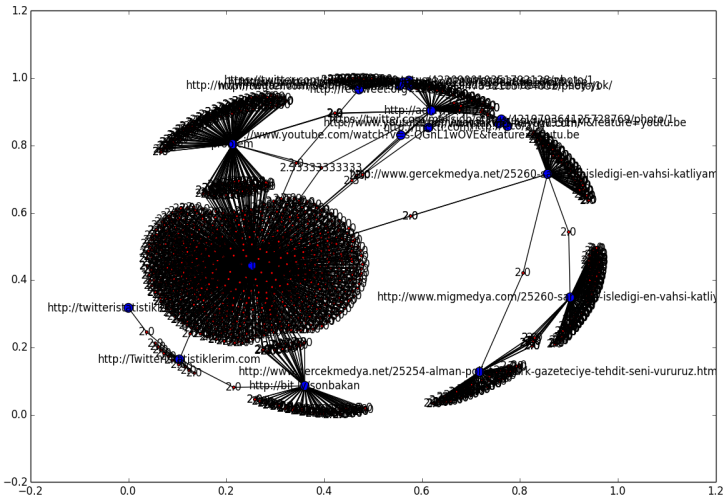
Some characters in the text have special functionality. Users can be directly mentioned with the '@' character. So the user 'JohnDoe' will be notified when the text '@JohnDoe' occurs in a tweet. Most significantly when the character '#' precedes a word in a tweet, then this word is considered a hashtag. Hashtags are something more than a simple pinpointing of significant words in a tweet. By collectively hashtagging certain words, a community can iconize certain words with meanings beyond the dictionary. It is a cultural phenomenon nowadays to associate certain hashtags with events and ideas.

Twitter communities, raise to become popular trends that hold a prominent place in Twitter's social impact. These trends can be location specific or worldwide and are a very compact representation of the current society's interests

Twitter is not immune to spam.



Source: http://xkcd.com/723/



A group of Spam Campaigns as plotted by our technique. The central blue node is the malicious link and the nodes connected to it are the users that have tweeted a message containing the specific

In fact, on average, 1% of total tweets contain spam with unsolicited advertisement or even worse with links to phishing or malware sites. Studies show that users are more vulnerable in Twitter spam than in traditional email spam. The probability a user on Twitter clicks a spam link is two orders of magnitude higher than in email spam.

Spammers use various techniques in order to force spam tweets to reach a maximum attendance. One of the most common one is to use unrelated trending topics. In our research, we study the extent in which trending topics are exploited by spammers.

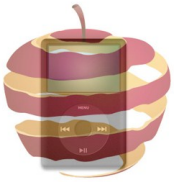
LinkedIn is for the people you know. Facebook is for the people you used to know. Twitter is for people you want to know (Unknown)

In order to achieve that we download a significant portion of all tweets that contain trending topics. Then for each user we extract various metrics than can give strong indications regarding the legitimacy of this user's tweets. These metrics can be for example different trending topics, number of links and tweets per day. Moreover, we know that spammers use collective techniques to spread spam. Fraudulent or exploited accounts can be utilized to send mass spam, otherwise known as 'spam campaigns'.

These campaigns can be represented as a network. The visualization of these networks can reveal commonalities between these campaigns that can help us build filtering mechanisms that act preemptively by flagging a tweet, or a link as potentially dangerous. The application of these defenses can make Twitter a safer place where these awesome tiny pieces of text come exclusively to amuse, inform us and interact with our peers.

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

Apple Skin Protects your iPod!



Document categorization is the task of dividing a set of documents into different groups such that all the documents in each group refer to the same entity in the real world. it's a daily routine in news agencies, where they categorize a collection of documents into different groups based on their topic to extract a summary out of each topic. Such categorization task requires to resolve the entity that is mentioned by each document and group documents based on those entities. However entity resolution is not a trivial task. Sometimes it's even difficult for humans to specify the real entity behind a document or a sentence. For example indicating the real entity mentioned by the title of this article (namely the "Skin" of the iPod) is difficult. Therefore we need to find a solution that can categorize documents without knowing the real entity they are talking about.

Which "Apple Skin:" is this sentence talking about? Is it the skin of the "Apple Fruit" or the skin made by the "Apple Company" that protects your iPod? Resolving such ambiguities requires extra information from documents talking about similar topic but containing less ambiguity. Extracting those documents from a data set, requires categorization of the documents in that data set into homogeneous groups such that all the documents in each group refer to the same entity in the real world. This categorization task is known as Cross-Document Coreference Resolution (CDCR)

To solve this problem we assume that similar documents are more likely to talk about same entity. In other words the referring expression of similar documents are more likely to refer to the same entity. Hence we categorize the set of documents based on their similarity instead of their topic. This type of categorization is known as clustering in which the number of clusters are not specified in advanced. Therefore one needs to find the best clustering in which all the documents in each cluster have maximum similarity to each other (maximum intra-cluster similarity), while documents from different clusters exhibit the minimum similarity to each other (minimum inter-cluster similarity).

This graph provides us with required information regarding the underlying similarity structure between different documents in the data set. Therefore the number of comparisons can be reduced to those of similar documents.

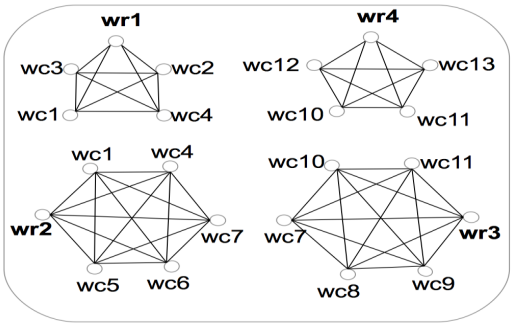


Figure 1: A fully connected graph called Clique for each document

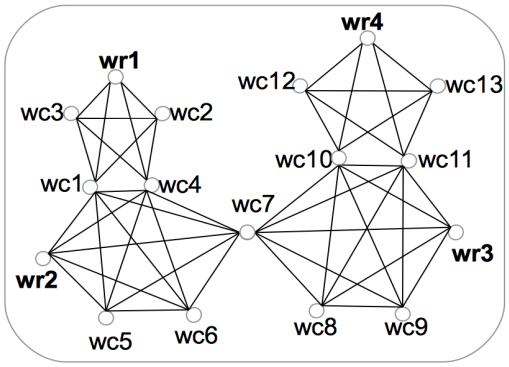


Figure 2: A fully connected graph after removing duplicates of common terms between different document

We proposed an innovative graph based modeling structure together with a diffusion based, node centric, clustering algorithm to solve this problem. Our graph based model converts each document into a graph. The graph contains the referring expression W_r together with its surrounding context words W_c. Vertices of the graph represent different words and the edges connecting them specify the co-occurrence of the words in each document. Hence we will create a fully connected graph called *Clique* for each document as **Figure 1** shows. This representation contains duplicates of common terms between different documents. To remove those duplicates and come up with a single, unique representation of the whole data set we connect cliques on their common words and create the graph represented in **Figure 2**.

In addition for every single similarity comparison we only compare documents on the similarity between their common terms instead of the whole unique term comparisons required on previous models. Looking at the graph, we can see that parts of the graph contain higher connectivity. Those are the parts representing documents with higher similarity. Therefore with a similar argumentation as before we can state that topological parts of the graph with higher inter connectivity are better candidates for co-reference relation between documents in those areas. In other words, having the graph structure we need to specify topological isolated or semi-isolated parts in the graph in order to find the best possible similarity clustering of the documents. We developed a diffusion based clustering algorithm to achieve such result.

Our algorithm initializes by assigning different color to each document (**Figure 3**). After the initialization round, the algorithm starts to diffuse colors into the graph applying specific diffusion strategies. Diffusion strategies are a set of rules that indicate cluster boundaries by increasing the concentration of different colors in different topological isolated or semi-isolated parts. For example, **Figure 4** shows the clustered results corresponding to the set of documents presented above.

The final coloring structure indicates that W_{r1} and W_{r2} are more likely to refer to the same entity in the real world as well as W_{r3} and W_{r4}. In other words we assume that all ambiguous words which are represented by the nodes of the same color are referring to the same concept/entity.

Our innovative solution outperformed the inference based probabilistic model developed by Google on a specific data set called John Smith with around 23% better results over the same evaluation measure.

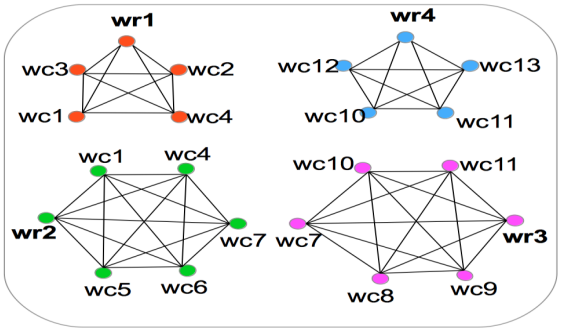


Figure 3: Assigning different color to each document

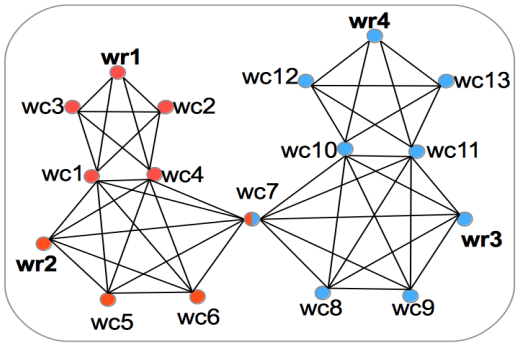


Figure 4: The clustered results corresponding to the set of documents

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

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

Newsletter Content Editor:
Paraskevi Fragopoulou
Foundation of Research and Technology-Hellas (FORTH)

Newsletter Designers:
Kalia Orphanou and Maria Poveda
Univerisy 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