iSocial

Issue 4, March 2017

ISOCIAL: DECENTRALIZED ONLINE SOCIAL NETWORKS

26 - 27 September 2016: "Decentralised OSN" Crete, Greece

Inside this Issue

TRENDING TOPICS ON DECENTRALIZED ONLINE SOCIAL NETWORKS

Middleware solutions for distributed pub-sub systems
Stefanos Antaris

Our research mostly focused on the design of novel P2P architectures for Distributed Pub/Sub Systems over DOSNs, that address both the problem of scalability and expressiveness that current centralized Pub/Sub systems confront.

See Page 4

Processing Big and Fast Data: Algorithms for Large-Scale Data Processing
M. Anis Nasir

The topics focus on designing algorithms for processing big and fast streams of data coming from various sources, e.g., event logs, sensor networks, network monitoring. We present algorithms to mitigate the challenges at both system and algorithm level. First, we focus on the load balancing problem that is caused by the skew in the data, heterogeneity in the cluster and skew in the request distribution. In doing so, we present three different algorithms to reduce the load imbalance across a distributed system. These algorithms are integrated into Apache Storm, which is an open source stream processing framework. Second, we present algorithms targeting problems in the domain of data mining and management. In particular, we present efficient algorithms for the top-k densest subgraph problem in the fully dynamic settings.

See Page 6

Managing online social network users's data
Alexandros Karakasidis

We developed efficient algorithms for problems related to graph partitioning, densest subgraph, and community detection under streaming settings. During the lifetime of the project, two internships were also completed

See Page 8

Location Based Access Control and Supernode Selection for online P2P video distribution
Giovanni Simoni

we have worked on how to efficiently select supernodes in a Peer-2-Peer (P2P) system deployed on the Internet. Second, we have contributed techniques that enables Location Based Access Control (LBAC) for a P2P video distribution platform.

See Page 9

Hive.js: Browser-Based Distributed Caching for Adaptive Video Streaming
Mikael Högqvist Tabor

In this work, we have investigated real-time P2P video streaming using WebRTC, a new W3C standard for browser to browser communication. Our initial results are encouraging, we have shown through a range of experiments that it is feasible to share live video streams between browsers with over 90% reduction in CDN traffic

See Page 10

Distributed Risk Assessment based on User Anomalous Behaviour in DOSNs
Laleh Naeimeh

We focused on identifying various security and privacy problems in OSNs, investigating the severity of these problems and the impact they impose on the user, and designing new systems and mechanisms for addressing these problems.

See Page 11

Security and Privacy Issues in OSNs
Panagiotis Ilia

Our research is focused in identifying various security and privacy problems in OSNs, investigating the severity of these problems and the impact they impose on the user and designing new systems and mechanism for addressing them.

See Page 12

Middleware Solutions for distributed Pub/Sub Systems

Online social networks (OSNs), such as Facebook, Twitter and Google+, have been transformed in a ubiquitous platform through where billions of people communicate and share user-generated content. Notifications constitute one of the primary way social users first learn about content that their social friends publish or their preferable sources (e.g. groups, pages) share. Obviously, Facebook, Twitter or Google+ are all examples able to disseminate a huge amount of notifications per day on billions of users. To achieve the desired scalability, OSNs deploy Publish/Subscribe (Pub/Sub) systems over distributed servers (brokers), where subscribers express their interest on specific events and are notified when an event that matches their interest occurs. Although large-scale Pub/Sub systems exist, they either typically require large company's resources to accommodate large-scale workloads of notifications or fail to support subscriptions with complex queries.

Our research mostly focused on the design of novel P2P architectures for Distributed Pub/Sub Systems over DOSNs, that address both the problem of scalability and expressiveness that current centralized Pub/Sub systems confront.

To overcome the aforementioned problems, Peer-to-Peer (P2P) networks provide a sound and highly scalable solution upon which distributed Pub/Sub systems could be build. However, current P2P networks are generic in the sense that they provide low-level communication abstractions, such as routing, of the physical network and they are not designed specifically for Pub/Sub systems. Therefore, integrating Pub/Sub systems with P2P networks is a challenging task that require several design decisions to be taken into consideration in order to perform scalability and reliability. In our research, we designed and build a Peer-to-Peer (P2P) architecture for Distributed Notification Systems over Decentralized Online Social Networks (DOSNs). Our architecture provides a novel ring topology of the P2P network and exploits the social graph to establish connections between peers. By exploiting the social graph in the P2P overlay network, we manage to reduce the number of overlay hops required for the communication between two users in the social network. Moreover, we experimentally showed that our approach reduces the total network latency between two socially connected users versus the state-of-the-art P2P overlay networks.

Moreover, we investigated the problem of supporting complex queries on the subscriptions of the social users. Thus, social users would be allowed to provide fine-grained subscriptions not only based on the social friends or groups that they want to connect, but also based on the content that their friends/groups produce. Specifically, we designed a novel Content-based Pub/Sub system, where social users express their interest following well-defined attribute/value pairs. Based on their interest, we organize social users with similar interests in the same group using a P2P gossip-based protocol. After forming K groups, where K is automatically identified by the clustering algorithm used, we apply a P2P ring topology using only the peers of the group. Thus, when a new post in the OSN is generated, it is automatically redirected to the group which contains the users mostly interested on and start propagating through the peers using the P2P ring topology. This research is still under development.

To sum up, our research mostly focused on the design of novel P2P architectures for Distributed Pub/Sub Systems over DOSNs, that address both the problem of scalability and expressiveness that current centralized Pub/Sub systems confront.

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

Processing Big and Fast Data: Algorithms for Large-Scale Data Processing

During the lifetime of the project, we dealt with several problems related to designing algorithms for big and fast input streams, e.g., event logs, sensor networks, network monitoring, using distributed stream processing systems. In doing so, we focused on handling both system and algorithm level problems. At the system level, we designed algorithms to solve problems related to network-aware load distribution, load balancing, fault tolerance and data management. Moreover, at the algorithm level we developed efficient algorithms for problems related to graph partitioning, densest subgraph, and community detection under streaming settings. During the lifetime of the project, two internships were also completed. The first one was at Yahoo Labs Barcelona and the second one was at Aalto University Finland. Further we provide the brief overview for each of the contributions.

We developed efficient algorithms for problems related to graph partitioning, densest subgraph, and community detection under streaming settings. During the lifetime of the project, two internships were also completed

Partitioning and Replication for Online Social Networks: Online Social Networks (OSNs) have been gaining tremendous growth and popularity in the last decade, as they have been attracting billions of users from all over the world. Such networks generate petabytes of data from the social interactions among their users and create many management and scalability challenges. OSN users share common interests and exhibit strong community structures, which create complex dependability patterns within OSN data, thus, make it difficult to partition and distribute in a data center environment. Existing solutions, such as, distributed databases, key-value stores and auto scaling services use random partitioning to distribute the data across a cluster, which breaks existing dependencies of the OSN data and may generate huge inter-server traffic. Therefore, there is a need for intelligent data allocation strategy that can reduce the network cost for various OSN operations. In this part of work, we developed a gossip-based partitioning and replication scheme that efficiently splits OSN data and distributes the data across a cluster. We achieved fault tolerance and data locality, for one-hop neighbors, through replication. Our main contribution was a social graph placement strategy that divides the social graph into predefined size partitions and periodically updates the partitions to place socially connected users together.

Load Balancing for Distributed Stream Processing Systems: Load Balancing problem deals with assigning the workload equally among set of machines in a distributed system. Such schemes enable achieving high throughput and low latencies. We studied the problem of load balancing in distributed stream processing engines, which is exacerbated in the presence of skew. We introduced PARTIAL KEY GROUPING (PKG), a new stream partitioning scheme that adapts the classical “power of two choices” to a distributed streaming setting by leveraging two novel techniques: key splitting and local load estimation. In so doing, it achieved better load balancing than key grouping while being more scalable than shuffle grouping. We tested PKG on several large datasets, both real-world and synthetic. Compared to standard hashing, PKG reduced the load imbalance by up to several orders of magnitude, and often achieves nearly perfect load balance.

Socially aware Distributed Hash Tables for Online Social Networks: Many decentralized online social networks (DOSNs) have been proposed due to an increase in awareness related to privacy and scalability issues in centralized social networks. Such decentralized networks transfer processing and storage functionalities from the service providers towards the end users. DOSNs require individualistic implementation for services, (i.e., search, information dissemination, storage, and publish/subscribe). However, many of these services mostly perform social queries, where OSN users are interested in accessing information of their friends.

In our work, we designed a socially-aware distributed hash table (DHTs) for efficient implementation of DOSNs. In particular, we proposed a gossip-based algorithm to place users in a DHT, while maximizing the social awareness among them.

Load Balancing for Distributed Stream Processing Systems under Skew: Carefully balancing load in distributed stream processing systems has a fundamental impact on execution latency and throughput. Load balancing is challenging because real-world workloads are skewed: some tuples in the stream are associated to keys, which are significantly more frequent than others. Skew is remarkably more problematic in large deployments: having more workers implies fewer keys per worker, so it becomes harder to "average out" the cost of hot keys with cold keys. We proposed a novel load balancing technique that uses a heavy hitter algorithm to efficiently identify the hottest keys in the stream. These hot keys are assigned to d ≥ 2 choices to ensure a balanced load, where d is tuned automatically to minimize the memory and computation cost of operator replication. The technique works online and does not require the use of routing tables. Our extensive evaluation showed that our technique can balance real-world workloads on large deployments, and improve throughput and latency by 150% and 60% respectively over the previous state-of-the-art when deployed on Apache Storm.

We provided a theoretical analysis of the proposed algorithm and showed empirically that the algorithm often generates denser subgraphs than state-of-the-art competitors.

Top-k Densest subgraphs in sliding windows graph streams: Given a large graph, the densest-subgraph problem asks to find a subgraph with maximum average degree. When considering the top-k version of this problem, a naive solution is to iteratively find the densest subgraph and remove it in each iteration. However, such a solution is impractical due to high processing cost. The problem is further complicated when dealing with dynamic graphs, since adding or removing an edge requires re-running the algorithm. In our work, we studied the top-k densest-subgraph problem in the sliding-window model and proposed an efficient fully-dynamic algorithm. The input of our algorithm consists of an edge stream, and the goal is to find the node-disjoint subgraphs that maximize the sum of their densities. In contrast to existing state-of-the-art solutions that require iterating over the entire graph upon any update, our algorithm profits from the observation that updates only affect a limited region of the graph. Therefore, the top-k densest subgraphs are maintained by only applying local updates. We provided a theoretical analysis of the proposed algorithm and showed empirically that the algorithm often generates denser subgraphs than state-of-the-art competitors. Experiments showed an improvement in efficiency of up to three to five orders of magnitude compared to state-of-the-art solutions.

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

Managing online social network users's data

My participation in the iSocial Marie Curie ITN was a unique, fruitful and multidimensional experience. I had the opportunity to be exposed to a wide variety of activities and perform research in the broader area of online social networks. Regarding my research, I have worked on managing online social network users' data, a subject that both aligns with my major research interest of data management and creates for me new research directions.

Efforts were directed towards the problem of users locating nearby social network friends, without the users, having to reveal their exact location either to the online social network, or to each other

The first problem I have worked on, regards to providing privacy to social network users. I have worked on managing online social network users' data, a subject that both aligns with my major research interest of data management and creates for me new research directions. The first problem I have worked on, regards to providing privacy to social network users. In specific, my efforts were directed towards the problem of users locating nearby social network friends, without the users, having to reveal their exact location either to the online social network, or to each other. Since users of such services are considered to be highly mobile, calculations for providing such services are continuous. To this end, the developed methodology is particularly lightweight, and proven to provide privacy, without the need of computationally expensive cryptographic methods, but exploiting randomization techniques.

The second problem I have worked on, focuses on fast resolution of entities resulting from streams of data. Currently, linking data between two sources involves the construction and update of an entire graph. In the case of social networks, where data form voluminous streams, expensive computations are required. To solve this problem, I worked on speeding up the process by avoiding the reconstruction of the entire graph and resulted into some very interesting initial theoretical results, regarding the order and the propagation of the updates.

Furthermore, I had the chance to receive training on proposal authoring. This way I acquired an additional skill which is going to be very useful for my future academic career. Furthermore, I attended scientific events and perform talks to a variety of audiences and performed a secondment which provided me with material for further future research.

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

Location Based Access Control and Supernode Selection for online P2P video distribution

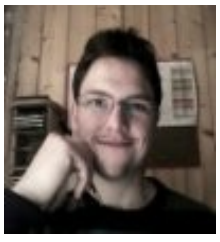
In recent years, the increasing demand for high quality multimedia content required a capacity boost of the communication infrastructures. Content Delivery Networks nowadays provide high throughput services, and support the delivery of large amounts of data in the Open Internet. In this work, we have addressed two key areas which also overlaps with distributed online social networks (DOSN). First, we have worked on how to efficiently select supernodes in a Peer-2-Peer (P2P) system deployed on the Internet. Second, we have contributed techniques that enables Location Based Access Control (LBAC) for a P2P video distribution platform.

First, we have worked on how to efficiently select supernodes in a Peer-2-Peer (P2P) system deployed on the Internet. Second, we have contributed techniques that enables Location Based Access Control (LBAC) for a P2P video distribution

In peer-to-peer applications deployed on the Internet, it is common to assign greater responsibility to supernodes, which are usually peers with high computational power, large amount of memory, or high network bandwidth capacity. In this paper, we describe a practical solution to the problem of supernode selection, that is the process of discovering the best peers in the network by some application-specific metric. We provide a distributed heuristic that allows to identify the best K nodes in the P2P overlay, by taking into consideration the realities of actual deployments, such as the presence of NATs. Our approach consists of an epidemic protocol which does not require new connections to be established, but rather relies on established connections, such as the ones provided by a NAT resilient peer sampling framework. We support our claims with a thorough evaluation of our solution in simulation and in a real deployment on thousands of consumer machines.

Location Based Access Control (LBAC) techniques allow security policies to account for the physical position of the user, enabling the enforcement of access rules based on his geographical location. In this work, we extend a P2P video distribution platform with LBAC enforcement management. The main target of our platform is minimizing latency and maintain an outstanding level of quality of user experience. For that, our application relies on retrieving content from multiple sources (peers and the CDN) simultaneously. We propose a LBAC architecture that is compatible with our application's approach and targets an enterprise scenario, where locations are physically well defined by the network accessibility, that is by the structure of the WiFi network. We evaluate our LBAC scheme by means of simulations, analyzing how the presence of a LBAC enforcing system affects the data delivery latency with respect to the standard scenario where LBAC restrictions are not enforced.

Giovanni Simoni
giovanni.simoni@hivestreaming.com
ESR iSocial Fellow
Hive Streaming AB, Sweden



Hive.js: Browser-Based Distributed Caching for Adaptive Video Streaming

In this work, we have investigated real-time P2P video streaming using WebRTC, a new W3C standard for browser communication. We have shown through a range of o browser experiments that it is feasible to share live video streams between browsers with over 90% reduction in CDN traffic.

Streaming video is one of the most important communication forms driven by Online Social Networks such as Youtube, Facebook and Twitter. It is estimated that by 2020, 82 percent of all consumer internet traffic will be video. In addition, 64 percent of that traffic will cross Content Distribution Networks (CDNs). Given the growth rate, current CDN technology may not be able to support the large amount of traffic in a cost-efficient manner. A promising alternative that has proven very effective for similar applications such as large scale file transfer is Peer-to-Peer (P2P) technology. With P2P, the viewers are directly connecting to each other without burdening the CDN with unnecessary traffic.

Up until now, adoption of P2P technology for video streaming has been hindered since, in addition to the web browser, a separate client-side plugin or software must be installed by the end-user. In this work, we have investigated real-time P2P video streaming using WebRTC, a new W3C standard for browser to browser communication. We have shown through a range of experiments that it is feasible to share live video streams between browsers with over 90% reduction in CDN traffic. In addition, the results of this study resulted in a new commercial project at Hive Streaming and is being deployed at multiple customers.

Mikael Högqvist Tabor
mikael.hogqvist@hivestreaming.com
ER iSocial Fellow
Hive Streaming AB, Sweden

Distributed Risk Assessment based on User Anomalous Behaviour in DOSNs

Decentralized Online Social Networks (DOSNs) described as internet-based applications that allow users to create a public or private profile, encourage them to share information and interact with other users and communicate with each other in the virtual environment.

Our experimental result on real Facebook data set showed that our proposed approach gained a comparable accuracy compared to the centralized scheme

In DOSNs, each user has some private information including his/her interactions and profile data and this local information cannot be moved to a central server or to other users in a raw form due to privacy issues.
On the other hand, once a user creates a profile and start to use DOSN, he/she establishes new relationships with strangers. In this case, if the user shares a huge amount of personal data with his contacts without specifying appropriate privacy settings, this can lead to security risks and can attract a variety of privacy and highly damaging attacks. These attackers create fake accounts or steal legitimate user accounts to forward spam and malware on online social networks.

In this paper, we propose a distributed two-phase risk estimation approach to detect risky users in DOSNs based on their behavioral patterns. Our goal is to analyze the behavior of users (interactions or activity patterns) in DOSNs by identifying those risky users who follow the behavioral patterns of attackers. The goal is to compare the behavioral patterns of users with other similar users in the network to find misbehaviors in a distributed manner.

More precisely, when the user behavior diverges from normal behavior, the user will be considered as risky. Our experimental result on real Facebook data set showed that our proposed approach gained a comparable accuracy compared to the centralized scheme.

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

Security and Privacy Issues in OSNs

The rapid emergence and increasing popularity of content sharing services and Online Social Networks (OSN) during the last years has radically changed the structure and the utility of the Internet. This explosive growth of OSNs has, in a sense, created the first digital generation consisting of people of all ages and backgrounds. People are creating their digital counterparts for interacting with others, for both recreational and professional reasons, and disclose a vast amount of data and sensitive personal information in an attempt to fully utilize these services. Inevitably, this behavior has raised many concerns regarding user privacy and the wide range of threats users expose themselves to, ranging from identity theft to monetary loss.

Focused on identifying various security and privacy problems in OSNs, investigating the severity of these problems and the impact they impose on the user, and designing new systems and mechanisms for addressing these problems

In order to facilitate these possible threats, we designed and built a framework for crawling a wide range of popular social and communication services, and for correlating accounts that belong to the same user across these different services. Our results, which serve as a cautionary tale, demonstrate how disjoint sets of information from multiple services can be associated and also, that large-scale, automated, personalized attacks are feasible. In another work towards this direction, we assess the Social Authentication mechanism employed by a popular OSN. We demonstrate that an adversary capable of collecting the publicly available photos of the victim and his friends can easily defeat the Social Authentication mechanism with two simple image comparison attacks, which can lead to account takeover. For this reason, we revisited Social Authentication and we built a system that retains the usability of the existing mechanism, while being robust to attacks employing image analysis software.

Moreover, we investigated the problem of privacy leakage from collective resources, that are associated to multiple users (e.g., group photos). The existing mechanisms employed by OSNs for controlling access to a photo cannot effectively handle cases where the interested parties have conflicting settings. In general, only the uploader is considered the owner of the photo and is granted full rights, whereas the people appearing in the photo are not considered co-owners and are not granted any rights. This leads to situations where the will of the publisher goes against the will of the depicted users, or the privacy settings of one user override those of another. Therefore, we design a new fine-grained access control mechanism that allows each individual depicted in a photo to specify its own privacy settings that control the visibility of the area that depicts its own face. In another work towards this direction, we also take into consideration the case of privacy leakage due to the service provider, and thus, we design a decentralized multi-party access control system that (i) allows collective privacy management and also (b) protects users from the service provider.

Overall, our research is focused in identifying various security and privacy problems in OSNs, investigating the severity of these problems and the impact they impose on the user and designing new systems and mechanism for addressing them, without impacting the usability of the system and affecting user experience.

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

Privacy Preserving Services for Social Networking sites - Towards engaging end-users

Decentralized Online Social Networks (DOSNs) have been investigated as an alternative transparent solution to protect users' privacy, and to avoid surveillance concerns resulting from the centralized model. The core idea is to give online social network (OSN) users full control over the management of their data, without having to delegate it to a central service provider. However, there are various technical challenges that a decentralized architecture imposes, and that have been limiting DOSNs to compete against the centralized model.

Our research resulted in design of two access control models that adopt approaches that have not yet been explored within OSNs: an aposteriori based, and a label based solutions.

One of the most prominent of these challenges is access control management. Most of the available proposals take a cryptographic approach. These solutions may provide strong security guarantees; however, they fail short at modeling the required fine-grain and dynamic access scenarios of OSNs, they introduce high operational costs, and they suffer to scale. Therefore, our main research question in this Ph.D. work is on what other alternatives could be exploited to offer flexibility in defining fine grain access controls in DOSNs, and that would scale better than encryption based techniques. Our research resulted in the design of two access control models that adopt approaches that have not yet been explored within OSNs: an aposteriori based, and a label based solutions.

On the path to achieving those solutions, and starting from an understanding of the access control process, we have found it necessary to dedicate efforts to the investigation of the problem of identity validation in OSNs as well. Identification comes as a critical step in ensuring the meaningfulness of any access control solution that could be put in place.

In OSNs, identification is more of a user responsibility than it is a system one. The looseness in obtaining an identity in an OSN (i.e., creating an account with a valid email address), makes it up to the user whether or not to trust the claimed identity on a profile. Thus far, there have been no proposals to assist users with such a task of reliable profiles identity estimation. Therefore, our additional contribution in this work came to put a block in this identified gap.

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

From Milgram to multiplexes: mutual greedy routing

My average intermediate degree of separation from everyone on Facebook is 3.39 [1]. This means that I probably know you through a chain of 3 to 4 intermediate individuals. That does not seem a lot, but how can one find these people? Facebook can calculate this number because they know the structure of the whole network: who is connected to whom, globally. But without this knowledge, how can I reach any individual?

The answer dates back to the origin of the famous "six degrees of separation". In 1967 Stanley Milgram conducted an experiment. Participants chosen as starting points of the study were provided with information about a target person to whom they had to send a letter. The results were quite surprising. On average, it took only five steps for the letter to reach the target. Five steps means six degrees of separation; an expression that has become famous since Milgram's experiment. This means that people can actually find the chains of intermediate individuals connecting everyone in a small number of steps. As Milgram puts it, "each intermediary moves the folder toward the target person. That is, a certain amount of information about the target person –his place of employment, place of residence, schooling, and so forth– is given to the starting subject, and it is on the basis of this information alone that he selects the next recipient" [2]. Just as people nowadays use Facebook and Twitter simultaneously, individuals in Milgram's study used information from multiple "domains" at the same time, e.g., the geographic location and occupation of an individual, to bring the message closer to its destination.

From a system-level perspective, a prospering future in the digital age comprised of a diverse digital landscape with interacting, decentralized architectures is possible, but so is the opposite

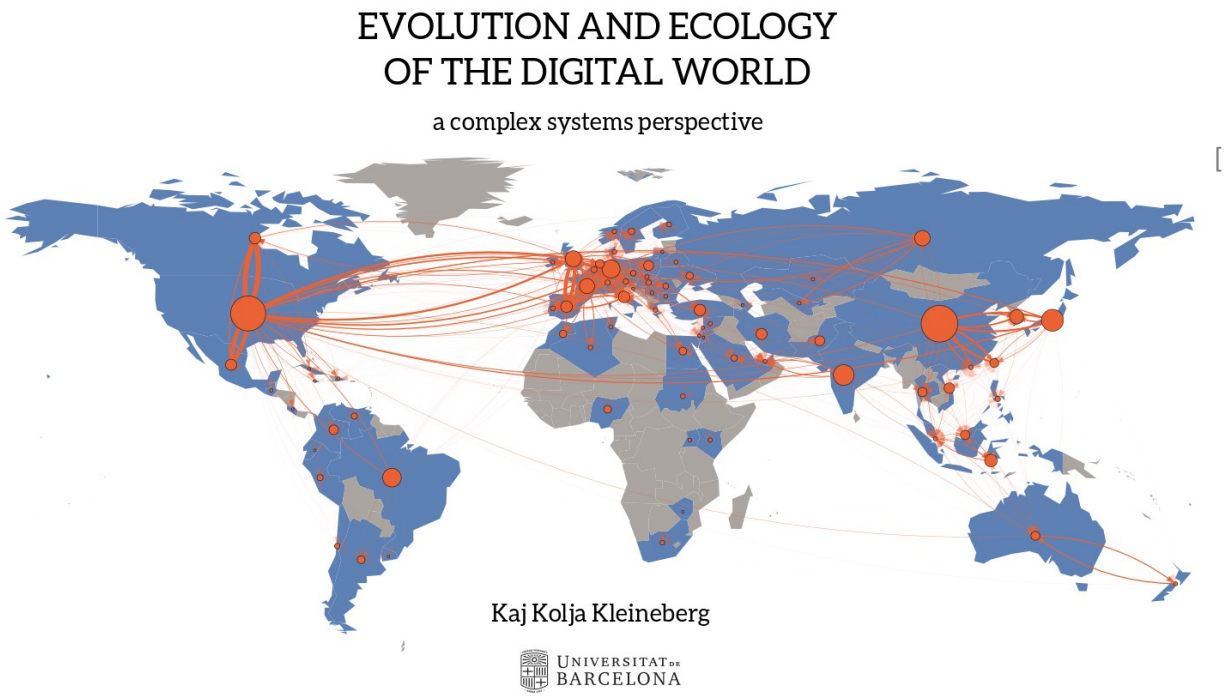
We can understand these different domains as different networks, like in the mentioned case of Facebook and Twitter. In these networks, individuals are represented as nodes. Individuals use all networks to disseminate information, which increases its spread significantly and can enhance the chance to reach other individuals. But, how can one reach any individual using both Facebook and Twitter, without the knowledge of who is connected to whom in the whole network? And, under which conditions is it better to use both networks compared to using only a single network?

We show that the key to answer these questions lies within hidden geometric relations between the different networks, which are present in many different real systems. Our findings reveal that these relations are optimal for reaching nodes in the system using several networks, in the same way as individuals reached the target in Milgram's experiment. In other words, if the discovered relations between the networks are present, individuals can be reached more efficiently if one uses several networks at the same time. Furthermore, we show that adding only a small number of other networks can increase the message transfer efficiency close to perfection. On the contrary, if the discovered geometric relations were absent, there would be no benefit of using multiple networks compared to using a single network.

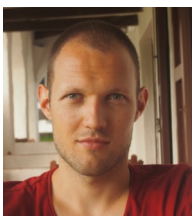
The discovered geometric relations across networks are a fundamental property of the organization of natural and human made systems. The universality of the presence of these relations suggests that there are some common underlying principles —like physical laws— that govern all of these systems. Our findings can have important applications in diverse domains, ranging from improving information transport and navigation or search in communication systems and online social networks, to understanding functional and structural networks in the human brain and deciphering their precise relationship(s), to predicting links among nodes (e.g., terrorists) in a specific network by knowing their connectivity in some other network.

:[1] https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/.
[2] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. Sociometry, 32 (4):425–443, 1969.

the related paper con be found here: http://www.nature.com/nphys/journal/v12/n11/full/nphys3812.html



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



Knowledge Extraction at Scale using Distributed Graph Algorithms

Extracting useful knowledge from large amount of unstructured data, constantly generated in *"Online Social Media"*, has become one of the main challenges in front of many companies in recent years. One example is disambiguation of artist names from a large amount of unstructured metadata at Spotify that became a significant barrier for delivering higher quality services to their customers. Another example is topic identification and tracking on Twitter that recently, become an important requirement for web intelligence companies like Gavagai (http://gavagai.se/) in their various applications on *"market research"* and *"opinion mining"*.

As the number of online social services increases the requirements and constraints for development of efficient and accurate methods for knowledge extraction from their data also increase. Scalability, accuracy and latency are examples of such requirements and constraints. For example, scalability is an important requirement for topic extraction on Twitter, where users publish around 500 million Tweets per day. Also, latency is a vital constraint in applications like auto driving vehicles, where sensor data must be collected, analyzed and the result must be fed back to the system, in orders of milliseconds, to provide accurate decision making and prevent accidents.

Throughout the duration of iSocial project we developed solutions to meet requirements and constraints and to overcome the limitations of knowledge extraction from large-scale data generated in online social media

Throughout the duration of iSocial project we developed solutions to meet requirements and constraints and to overcome the limitations of knowledge extraction from large-scale data generated in online social media. We designed and implemented various algorithms for (i) semi-supervised multiple disambiguation of ambiguous words in large corpuses of natural language text, (ii) large-scale topic identification on Twitter using dense and weighted graph partitioning and (iii) geo-location Identification over Twitter short messages using context, time and social-network information. To develop these solutions first, we applied various preprocessing approaches like exploratory data analysis, sampling and hypothesis testing for understanding the data and sketching the problem. Then, we projected the problem into a graph representation using information collected in previous step.

Finally, we implemented a distributed algorithm that runs on top of the graph representation to extract the intended knowledge. Our first solution outperformed the best contemporary solutions with around 10% improvement in terms of accuracy. The second solution achieved similar or better quality on topic modeling, but 10 times faster than the state-of-the-art approaches.

Finally, on Twitter geo-location identification, our solution shrunk the average prediction error to less than 1 km on 20% of the training data, which significantly outperformed current best approaches.

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

Graph-based Analytics for Decentralized Online Social Network

My research is dedicated to integrate graph analytics with machine learning for the new trend in Online Social Networks, which is Decentralized Online Social Networks (DOSNs). DOSNs require large-scale distributed machine learning algorithms that would be highly scalable to the huge number of users and the large amount of social data streams.

Furthermore, such algorithms need to evolve with respect to temporal, spatial, and context factors of users' interactions. However, existing centralized graph-based analytics are not realistically scalable due to the huge number of users. Therefore, the core of my work is to provide massively parallel graph-based algorithms that suitably fit DOSNs and to eliminate the need of single centralized aggregation point. Additionally, existing machine learning frameworks are data-centric, such that they rely heavily on various data collection, extraction, and analysis technologies. Yet, social networks are categorized as complex systems, capturing the fact that it is difficult to derive their collective behavior by extracting knowledge from available data without considering the intricate network that encodes the interactions between users. Accordingly, the main objective of my work is to investigate the best approaches to combine network and data analytics. Specifically, this integration extracts knowledge that encodes both of the topological and behavioral interactions between users as well as patterns extracted from the shared data.

The core of my work is to provide massively parallel graph-based algorithms that suitably fit DOSNs and to eliminate the need of single centralized aggregation point

So as, by integrating graph analytics with machine learning, we will be able to analyze autonomous data sources as well as users interactions in distributed and decentralized way.

Specifically, we have provided different algorithms that can be used to minimize risk and insecurity in DOSNs by developing privacy preserving trust management schemes for identity validation, anomaly detection and spam filtering. Each tool we develop is tested on real data and its value is judged by the insights it offers about a system's properties and behavior.

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

Multi-perspective analysis of OSNs: Spam detection, Network evolution and Content Analysis

Over the last years, Online Social Networks (OSNs) have been prevailed as the major medium for communication, news dissemination and social interaction for a big proportion of the population. It is estimated that as of 2016, approximately half of the world population has used at least one OSN within the past year. Today OSNs are in the edge of Computer Science and Internet technology since they have to meet very demanding expectations regarding efficiency, security and user interaction. OSNs is a phenomenon of our time with extreme significance and it has attracted the interest of many research communities that try to delve into the societal effects, uncover potential dangers and reveal hidden patterns of the social graph.

I revealed a vulnerability of Google search page which was exploited in order to propagate spam links in Twitter

During my PhD I explored all these three aforementioned areas by applying advanced methods from Social Network sampling, Machine Learning, Data Mining, Graph Theory, Modeling and Natural Language Processing. The major findings of my research per major research area are the following: In the area of vulnerability detection and attack prevention, I revealed a vulnerability of Google search page which was exploited in order to propagate spam links in Twitter. In the area of social graph analysis, I proposed a computationally inexpensive method to study the evolution of large (~100 million nodes) social graphs and locate periods of fast or slow growth. Finally in the area of societal effects I assembled methods to qualitatively assess the political inclinations and sentiments of Greek Twitter users towards certain entities in times of political upheaval.

As OSNs become more technologically advanced with high marketing potentials we expect that their effect on our social and private life will be more prevalent. For this reason, studies that shed light to the dynamics, content and security of OSNs like these included in my Thesis are not only exciting in a research perspective but also of extreme importance for our society.

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

Dynamical properties of the herding voter model with and without noise

We consider dynamics of a complex system modeling the behaviour of a large heterogeneous population. Starting from compartmentalised voter model with two separate time scales for ”fast” and ”slow” interactions, we augment it with introducing a separate stochastic process, responsible for independent opinion formation. This simple modification leads to emergence of rich variety of behavioural modes can be interpreted as herding behaviour/opinion leadership, information cascades, consensus oscillations, opinion polarisation etc. We analyse this new model which shows an interesting phase diagram, with a purely diffusive phase, a herding (or two-states) phase, and a mixed phases where both behaviours are possible.

The voter model with separate time scales and noise responsible for independent opinion formation shows the emergence of rich variety of behavioural modes with distinct phase separations in configuration space, that can be interpreted as herding behaviour/opinion leadership, information cascades, consensus oscillations, opinion polarisation etc.)

We show that, in the herding phase, a group of agents become effectively the leaders of the dynamics while the rest of the population follow blindly their opinion.

Interestingly, in some cases such herding dynamics accelerates the time to consensus, which then become size independent or, on the contrary, makes the consensus nearly impossible. These new behaviours have important consequences when an external noise is added to the system that makes consensus (absorbing) states to disappear. Thus, introducing spontaneous formation of independent opinion allows for modes where opinion distribution fluctuates indefinitely, instead of always converging to one of possible opinions like in classic voter models.

As we have seen, the addition of small variations to the classical voter model increases the range of possible dynamical behaviours dramatically. Heterogeneity in the activity rates of agents, coupled with a preference choice for active agents, induce the emergence of collective leadership in a fraction of the population while the rest simply follow the opinion of the leading group. This has important consequences for the global consensus time, which now range from being a constant value independent of the system size to an exponential function of the system size, in stark contrast to the standard voter model. The addition of intrinsic noise to the previous model makes its dynamics even richer, with the emergence of four distinct dynamical phases with distinct behaviour separations. Describing and identifying these phases has immediate applications in behavioural analysis of financial markets, opinion dissemination and other complex social modeling problems.

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

Online Social Network Evolution: Revisiting the Twitter Graph

Hariton Efstathiades (PhD candidate) from the Department of Computer Science of the University of Cyprus, received the Best student paper award for his work on Twitter graph analysis entitled ”Online Social Network Evolution: Revisiting the Twitter Graph” presented at the 2016 international conference on Big Data that took place in Washington, DC, USA on the 7th of December 2016. The paper was co-authored with Demetris Antoniades, George Pallis, Marios D. Dikaiakos from the University of Cyprus and Zoltán Szlávik, Robert-Jan Sips from IBM (Netherlands).

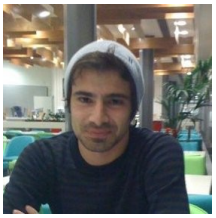
Our study retrieved 34.66M users connected by 2.06B social connections. Performing a comprehensive study of the 2009 and 2015 social graph snapshots and presenting the results regarding various metrics in the topology of the social graph

Twitter is one of the most popular Online Social Network (OSN) to time. It first appeared in 2006 and has been receiving growing attention ever since. As of 2015, the platform has more than 500 million users, out of which 316 million are considered active, i.e. users who log into the service at least once a month. Twitter allows its users to publish short messages, 140 characters long (including videos, pictures and URLs), in order to communicate their ideas, products, emotional state with their followers. Over the years Twitter has been used in a variety of different situations, e.g. allowing protesters to communicate over the Arab Spring. The extensive usage of Twitter enables researchers to analyze the generated information for several applications such as event detection, user location analysis, health care, recommendation and early warning systems, temporal trends and information diffusion.

Hariton’s work examines the Twitter network as it appeared in 2009 in the first comprehensive study of Twitter by Kwak et al. What is twitter, a social network or a news media? And, re-collects the users full characteristics as of late 2015.

In total the study by H. Efstathiades retrieved 34.66M users connected by 2.06B social connections. Performing a comprehensive study of the 2009 and 2015 social graph snapshots and presenting the results regarding various metrics in the topology of the social graph.

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



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

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

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

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

Find us on facebook

https://www.facebook.com/ISocialMarieCurieITN