Community structure and evolution analysis of OSN interactions around real-world social phenomena

Conference Paper · September 2013

DOI: 10.1145/2491845.2491849

CITATIONS
5

READS
243

3 authors:

Konstantinos Konstantinidis
Anatolia College - American College of Thessaloniki
22 PUBLICATIONS **243** CITATIONS
SEE PROFILE

Symeon Papadopoulos
The Centre for Research and Technology, Hellas
260 PUBLICATIONS **4,909** CITATIONS
SEE PROFILE

Ioannis (Yiannis) Kompatsiaris
The Centre for Research and Technology, Hellas
1,038 PUBLICATIONS **14,469** CITATIONS
SEE PROFILE

Community Structure and Evolution Analysis of OSN interactions around Real-World Social Phenomena

Konstantinos Konstantinidis, Symeon Papadopoulos, and Yiannis Kompatsiaris
Centre for Research and Technology Hellas, Information Technologies Institute
6th km Charilaou-Thermi Road
GR-57001 Thessaloniki,Greece
{konkonst,papadop,ikom}@iti.gr

ABSTRACT

In recent years, Online Social Networks (OSNs) have been widely adopted by people around the globe as a means of real-time communication and opinion expression. As a result, most real-world events and phenomena are actively discussed online through OSNs such as Twitter and Facebook. However, the scale and variety of such discussions often hampers their objective analysis, e.g. by focusing on specific messages and ignoring the overall picture of a phenomenon. To this end, this paper presents an analysis framework to assist the study of trends, events and interactions performed between online communities. The framework utilizes an adaptive dynamic community detection technique based on the Louvain method to study the evolution, overlap and cross-community dynamics in irregular, dynamically selected graph snapshots. We apply the proposed framework on a Twitter dataset collected by monitoring discussions around tweets containing extreme right political vocabulary, including messages around the Greek Golden Dawn party. The proposed analysis enables the extraction of new insights with respect to influential user accounts, topics of discussion and emerging trends, which could assist the work of journalists, social and political analysis scientists, but also highlights the limitations of existing analysis methods and poses new research questions.

Categories and Subject Descriptors

H.3.4 [**Social Networking**]: Online Social Network Community Analysis

General Terms

Algorithms, Experimentation

Keywords

Online Social Networks; Community Evolution Detection

PCI 2013, September 19 - 21 2013, Thessaloniki, Greece

`http://dx.doi.org/10.1145/2491845.2491849`

1. INTRODUCTION

OSN platforms such as Twitter and Facebook have become influential means of spreading trending news and ideas on emerging social phenomena. Such networks combined with advanced statistical tools are often seen as the best sources of real-time information about global events [12].

Twitter is a major OSN platform, with more than 200 million active users[1]. The range of content in tweets is extremely variable, ranging from what a person had for breakfast, to Barack Obama commenting on budget cuts. In 2009, Twitter played an important role in the Iranian election [4]. Due to the restrictions enforced by the Iranian government to bar journalists from "unauthorized" demonstrations, there was a significant lack of sufficient conventional news coverage. This actually drove the Iranian tech-savvy people to take up the important task of informing and communicating with the general public by disseminating news and events concerning the election via the Twitter network.

In fact this is only one out of countless examples regarding Twitter's use in political agendas, disputes and emerging social phenomena. Despite the fact that Twitter has turned into a massive dynamic data source for political analysis, the vast amounts of information shared through it cannot be accessed or made use of unless this information is somehow organized. Thus, appropriate means of filtering, sorting and summarization are necessary to support efficient browsing, searching and gaining an overall view of online discussions. Existing information browsing facilities, such as simple text queries typically result in immense amounts of posts rendering the inquirer clueless with respect to the online topics of discussion. In addition, important questions related to the origin and spread of online messages, as well as the dynamics of interactions among online users remain unanswered.

To this end, we propose a graph-based community evolution framework used to extract valuable information regarding events, discussions and other social phenomena with respect to four different factors. Specifically, we consider the keywords around a specific topic, the user activity in the form of mentioning posts, the communities to which the users belong, and most importantly, the evolution of these communities. The proposed analysis is carried out in four major steps: first, the Twitter API is used to extract mentioning messages that contain both an interaction between users and a keyword of interest. Second, a timeslot sequence is spanned in accordance with the users' level of activity. Next, graph snapshots are created based on user interactions and communities of highly interacting users are extracted.

[1]As "tweeted" by its own official account on Dec 18th, 2012

In the last step, the community evolution is studied to extract insights with respect to the interactions' lifecycle.

Two common issues in many dynamic data clustering procedures include the time interval at which each snapshot is formed and the similarity metric used to compare two parts of a potentially evolving community [6]. In this paper, we propose a novel activity-based adaptive time interval and a population dependent similarity metric. Additionally, we present an analysis of a Twitter dataset consisting of 880K mention messages exchanged between 857K Twitter accounts focused on the interactions performed between global communities associated with extreme right political vocabulary, including tweets around the recently popular Greek Golden Dawn (GGD) party.

Valuable insights are extracted from the analysis with respect to influential communities, topics of discussion and emerging events, which could assist the work of journalists, social and political analysis scientists. The typical journalistic Twitter analysis method regards the monitoring of specific users (news hounds), main events and mainstream trends. Although the latter usually provides the journalists with the most important trends and events, other less significant news could be retrieved by further analysis of regular users and their communities. An example of an event being sufficiently covered and heavily monitored using conventional journalistic methods could regard a national political party meeting. On the other hand, a minor event such as the speech of a municipality representative never makes the mainstream news but could for some reason pose an interest to the locals and as an extension to regional journalists. Additionally, the framework could be used to discover other minor happenings that took place in the national meeting but were overshadowed by the main event.

In this paper, a total number of 89K communities was extracted from the interaction network of 857K users including over 1M pairwise interactions and spanning a period of 32 days. Moreover, out of the full set of communities we analyzed 7K evolving communities that provided information from less dominant though persistent users. Through the study of the evolution of these communities a variety of interesting events emerged, information and opinions which do not usually register high on the retrieval rank of algorithms seeking for significant events, popular trends or celebrity opinions, thus demonstrating the value of the proposed framework for long tail news discovery. In addition, small Greek communities discussing the effect and sharing thoughts regarding the actions and announcements of the GGD party were also discovered.

The rest of the paper is organized as follows: Section 2 reviews relevant research efforts. Section 3 presents the details of the proposed framework. The experimental study along with the discussion are provided in Section 4 and the concluding remarks are made in the final section.

2. RELATED WORK

Mining OSN interactions is a topic that has attracted considerable interest in recent years. One of the most recent attempts comes from McKelvey et al. [14] who presented the Truthy system for collecting and analyzing political discourse on Twitter, providing real-time, interactive visualizations of information diffusion processes. They created interfaces containing several key analytical components. These elements include an interactive layout of the communication network shared among the most retweeted users of a meme and detailed user-level metrics on activity volume, sentiment, inferred ideology, language, and communication channel choices.

TwitInfo is another system that provides network analysis and visualizations of Twitter data. Its content is collected by automatically identifying "bursts" of tweets [13]. After calculating the top tweeted URLs in each burst, it plots each tweet on a map, colored according to sentiment. TwitInfo focuses on specific memes, identified by the researchers, and is thus limited in cases when arbitrary topics are of interest. Both of the aforementioned frameworks present an abundance of statistics for individual users but contrary to our method, they do not take into account the communities created by these users or the evolution of these communities.

Greene et al. presented a method [10] in which they use regular fortnight time intervals to sample a mobile phone network in a two month period and extract the communities created between the users of the network. Although the network selected is quite large (4M users) and the method is also very fast (1M nodes in 85 seconds); the system was created in order to be applied on a mobile phone network which renders it quite different to the network studied in this paper. The collected data lack the topic of discussion and the content of the messages between users, so there is no way to discover the reason for which a community was transformed or the effect that the transformation really had on the topic of that community. Moreover, in contrast to our method, the sampling intervals are regular (a fortnight), meaning that it does not take user activity into consideration.

Another interesting dynamic community detection method used to extract trends was introduced by Cazabet et al. [5]. They create an evolving network of terms, which is an abstraction of the complete network, and then apply a dynamic community detection algorithm on this evolving network in order to discover emerging trends. Although the algorithm is very effective for locating trends, it does not consider the interactions made between various users or the evolution of the communities.

To enhance the effectiveness and the information extraction capabilities of the above techniques and to enable their usage across a variety of networks, we propose the following framework that takes into consideration the temporal activity levels of user interactions and the corresponding community structure evolution.

3. OSN ANALYSIS FRAMEWORK

OSN applications comprise a large number of users that can be associated to each other through numerous types of interactions. Graphs provide an elegant representation of data, containing the users as their vertices and the interactions (e.g. mentions) among them as edges.

3.1 Notation

In this paper, we employ the standard graph notation $G = (V, E, w)$, where G stands for the whole network; V stands for the set of all vertices and E for the set of all edges. In particular, we use lowercase letters (x) to represent scalars, bold lowercase letters ($\mathbf{x}$) to represent vectors, and uppercase letters (X) to represent matrices. A subscript n on a variable (X_n) indicates the value of that variable at discrete time n. We use a snapshot graph to model interactions at a discrete time interval n. In G_n, each node $v_i \in V_n$

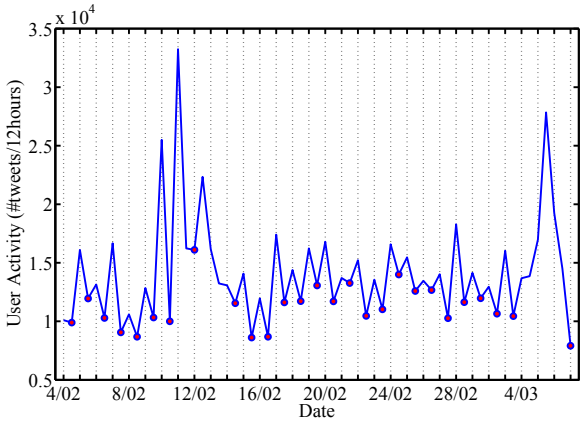

Figure 1: User activity on a 12-hour sampling rate.

represents a user and each edge $e_{ij} \in E_n$ is associated with a directed weight w_{ij} corresponding to the frequency of mentions between v_i and v_j. The interaction history is represented by a sequence of graph snapshots $\langle G_1, G_2, ..., G_n, ... \rangle$.

3.2 Framework Description

This section describes the proposed framework in three parts: interaction data discretization, community detection and community evolution detection and analysis.

3.2.1 Interaction Data Discretization

Taking into consideration that Twitter is a global network, selecting a fixed sampling rate to extract the desired graph snapshots would lead to inaccurately segmented communities. This is due to the fact that online discussions around events and trending topics do not often follow a strict circadian rhythm, since breaking news can happen at any time (leading to bursts of online discussions). This is further exacerbated by the wide use of mobile devices, which enable users to conduct their online discussions independent of their location and therefore making their timing less predictable.

In order to enhance the crude quantization of fixed time segmentation, we introduce an activity-based discretization technique. When the level of user activity reaches a local minimum as indicated by the red dots in Figure 1, the system introduces a break and a snapshot of the interactions between the users is extracted. The local minima are located by using the discrete first derivative of the user activity vector. As shown in Figure 1, an abundance of points at specific time samples are skipped since at those points the activity of the network either continues to drop signifying that there is no new event or discussion attracting the users' attention, or continues to rise thus indicating that the users' interest on some particular subject has not been lost.

3.2.2 Community Detection

Given a social network, a community can be defined as a subgraph comprising a set $V_{comm} \subseteq V$ of users that are typically associated through a common element of interest. This element can be as varied as a topic, a real-world person, a place, an event, an activity or a cause [16]. We expect to discover such communities by analyzing mention networks on Twitter. There is considerable work on the topic and a host of different community detection approaches appear in literature [8, 16]. Due to the nature of Twitter mention networks, notably their sparsity and size, in this paper we selected a community quality optimization method, more specifically the iterative heuristic scheme of the Louvain method [3]. The method is a greedy optimization method that attempts to optimize the modularity of a network partition [7]. The optimization is performed in two steps; initially, the method seeks small communities by optimizing the local modularity. In the second step, it sums up all the nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are iteratively repeated until a maximum of modularity is attained and a hierarchy of communities is produced.

The method has gained wide acceptance by the respective research community due to its computational efficiency and high precision, as well as the availability of an efficient reference implementation. Hence, we opted for its use. In the future, it may be interesting to investigate the sensitivity of the analysis results with respect to the selection of the community detection method.

3.2.3 Community Evolution Detection

The problem of finding communities in static graphs has concerned researchers for several years. However, the highly dynamic nature of OSNs has moved the spotlight to the subject of dynamic graph analysis [15, 1, 18, 9].

In this paper, we represent a dynamic network as a sequence of graph snapshots $\langle G_1, G_2, ..., G_n, ... \rangle$. The objective is to identify the communities $C = \{C_{1n}, C_{2n}, ..., C_{kn}\}$ that are present in the network across n timeslots. Each time-evolving community T_i is represented by a timeline of the communities it comprises.

An example of the most frequent conditions that communities might experience is presented in Figure 2: birth, death, irregular occurrences, merging and splitting, as well as growth and decay that register when a significant percentage of the community population is affected. In this example, the behavior of six dynamic communities is studied over a period of 3 timeslots. Dynamic community T_1 results from a previous timeslot and splits up into two new dynamic communities; the largest of the two continues as T_1 and the smaller one as T_7. This means that for some reason the members in T_1 at timeslot $n-1$ were split up into two separate smaller groups, which also explains the change in size. In our case it could be that a large group of supporters and opposers of the extreme right party engaged in conversation during $n-1$ but split up and are not cross mentioning each other in n and $n+1$. Moreover, the second group that formed the T_7 dynamic community, continued its decaying activity for one more timeslot and then stopped thus signifying the users' loss of interest in the discussion.

An opposite example is that of T_2 and T_3 in which two communities started up small but evolved through a merger into one very strong, large community that continues on to $n+2$. In this case it could be that two different groups of people witnessed the same event and began conversing on it separately. As time went by, connections were made between the two groups and in the n timeslot they finally merged into one. Actually, the community continued to grow as shown on the $n+1$ timeslot. T_4 and T_6 were both created (community birth) in $n-1$ and both disappeared in n differentiating in that T_4 reappears in $n+1$ (irregular occurrence) while T_6 does not. This is the main reason why a timeslot delay is

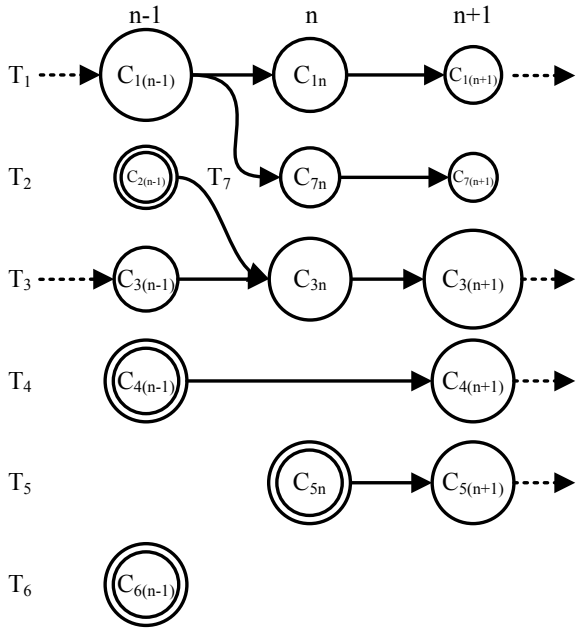

Figure 2: Example of five dynamic communities tracked over three timeslots, featuring (from T_1 to T_6) splitting, merging, skipping timeslots, birth (concentric circles) and death events. T_1 and T_3 continue from timeslot $n-2$, C_{5n} is born at n and continues, and $C_{6(n-1)}$ is born at timeslot $n-1$ and is discontinued just as $C_{2(n+1)}$.

introduced in the system as will be described later in this Section; searching for similar communities strictly in the immediate precedent timeslot would result in missing such possible re-occurrences.

To study the various lifecycle stages of a community, the main challenge pertains to the computational process used to identify and follow the evolution of any given community. On the one hand, it should be able to effectively map every community to its corresponding timeline, and on the other hand it should be as less of a computational burden as possible to be applicable to massive networks such as the ones induced by Twitter interactions.

However, community matching techniques presume a zero-to-one or one-to-one mapping between users in two communities, thus not supporting the identification of the above conditions in the lifecycle of a dynamic community. In order to overcome this predicament, we extend a recently proposed heuristic [10] relying on a user-defined threshold to determine the matching between communities across different timeslots. We propose an adaptive threshold technique, which alters the decision threshold exponentially with respect to the size of the communities at each timeslot.

More specifically, the algorithm steps are presented as follows. Initially, the first set of communities $\{C_{11}, C_{21}, ..., C_{k1}\}$ (i.e. the first snapshot) is extracted by applying the Louvain community detection algorithm [3] to the G_1 graph. A dynamic community marker T_i is assigned to each community from this snapshot. Next, the second set of communities is extracted from the G_2 graph and a matching process is performed between all the community combinations from the two consecutive snapshots in order to determine any possible evolution from the first snapshot to the next. The dynamic communities $T_{(1,2,...,i)}$ are then updated based on that evolution. For example, if C_{a1} does not appear in the second snapshot, T_a is not updated; a split is registered if the community appears twice in the new timeslot, and a merger marker is assigned if two or more communities seem to have merged into one. In fact, to avoid potential false positives of community deaths, a waiting time of approximately two days (i.e. four snapshots) is provided. If the evolution of a community is not detected in the last timeslot, the system queries the two previous ones in a "last come, first served" order. If no matching community is found, the community is considered dead. The evolution detection procedure is repeated until all graphs have been processed.

In order to determine the matching between communities, the Jaccard coefficient is employed [11]. Following comparative preliminary results between the Jaccard and the Sorensen index (dice coefficient) [17], the former was selected due to its efficiency as it proved to perform better. The similarity between a pair of consecutive communities C_{in} and $C_{(i(n-td)))}$ is calculated by use of the following formula, where timeslot delay $td \in [1, 3]$:

$$J\left(C_{in}, C_{i(n-td)}\right) = \frac{\left|C_{in} \cap C_{i(n-td)}\right|}{\left|C_{in} \cup C_{i(n-td)}\right|} \quad (1)$$

If the similarity exceeds a matching threshold ϕ, the pair is matched and C_{in} is added to the timeline for the dynamic community T_i. However, ϕ is not a constant threshold as in [10]. It varies in accordance with the community size. Since communities may range from a size of three to any significantly large number, it seems inappropriate to use a single threshold. For example, for a community of three and another of 100 users; selecting a constant threshold of 0.3 would mean that in order for the first community to be considered evolving it only has to match a single user in a relatively small community of a different timeslot. On the other hand, for the large community to be considered evolving it would require a match of at least 30 users. In order to overcome this predicament, we propose a threshold $\phi \in [0.1, 0.7]$ which is inversely proportional to the population, in such a way that the largest community of every timeslot only has to match 10%, while any communities of three have to match 70% and for the communities in-between the threshold follows an exponential tail. The limits of ϕ were heuristically selected after careful consideration of our experimental results and those reported in [10].

4. EXPERIMENTAL STUDY

Despite the proliferation of dynamic community detection methods, there is still a lack of benchmarked ground truth datasets that we could use to test our framework. Instead, the results presented in this paper were attained by applying our framework on a custom Twitter interaction network dataset that we make publicly available[2]. This real world dataset is a collection of mentioning posts acquired by a crawler that collects tweets containing any of 40 Greek or English terms, associated with the extreme right movement worldwide and in Greece (Table 1). The crawler ran over a period of 32 days, extracting 880K messages containing

[2]http://www.socialsensor.eu/results/datasets/82-twitter-dataset-pci2013

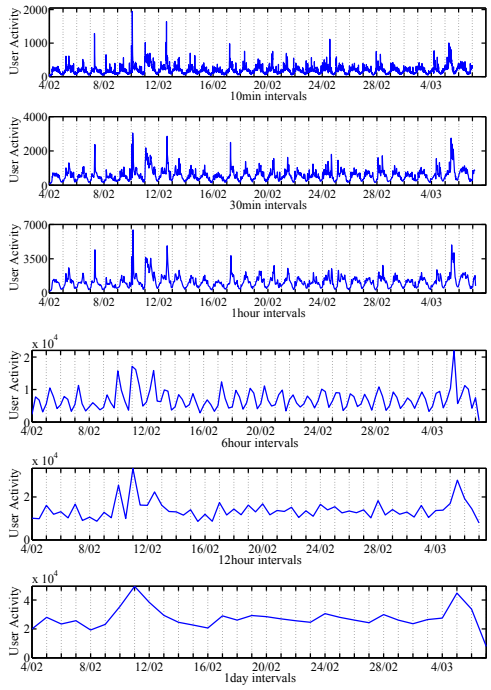

Figure 3: User activity over six time granularities: 10- and 30-minute, 1-, 6- and 12-hour, and 1-day).

mentions, 857K unique users and over 1M edges. The information we sought pertained to the various communities created between people who interact via mentions using extreme right vocabulary, people who are influenced by these communities and the various connections that exist, if any, between the respective Greek and foreign communities.

The network data was preprocessed as follows. Initially, all interaction data was transformed into weighted and directed adjacency matrices. The data was processed in accordance with the discretization technique described in section 3.2.1 resulting in a sequence of activity-based snapshots. Figure 3 displays the activity of the network on the basis of six different time granularities (1/6, 1/2, 1, 6, 12 and 24 hour granularities). The 12-hour time granularity resulted in the most discrete change in activity and thus was selected in order to create the sequential graph snapshots.

Next, a small number of users displaying an unusual high degree of self-loops (self-mentions) were removed as they correspond to accounts who are trying to manipulate their influence score on Twitter. In fact these accounts pose almost no interest to the communities since they mostly receive very few mentions. A good example is that of the dataset's second most active user who displays an activity

Table 1: Hashtag and keyword samples describing the far right movement for a) Greece and b) globally.

Greek		Global	
Hashtags	Keywords	Hashtags	Keywords
	Michaloliakos		nazi
#Xryshaygh	Kasidiaris	#nazi	far right
#GoldenDawn	golden dawn	#extremeright	extreme right
#Kasidiaris	xrysh aygh	#farright	Hitler
	illegal immigrants		Swastica

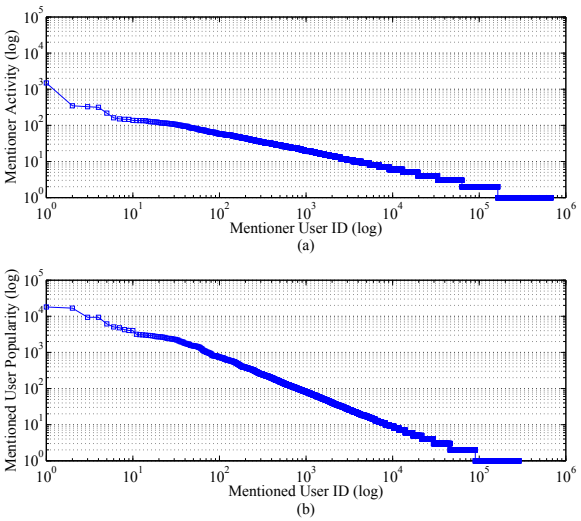

Figure 4: Distributions (in log scale) of numbers of (a) received and (b) posted mentions.

of approximately 260 posts. Although s/he seems to be a very active user on the basis of the posted mentions, almost all of them are self-loops. In addition, the mentions s/he receives from other people are far too few for the user to be considered influential.

Figure 4 depicts the Twitter user popularity in relation to the received mentions. It is interesting to note that the popularity distribution of mentioned users is far wider than the distribution of users posting the mentions. This indicates that there are particularly popular users who attract the attention of the rest of the community when discussions about extreme right topics, events and the behavior of specific people are concerned. As expected, both groups follow a long tail distribution, in which the mentioned users' coefficient is much steeper than the mentioners' one.

As described in subsection 3.2.2, for community detection we use the modularity optimization algorithm introduced by Blondel et al. [3]. Figure 5 presents a graph containing the modularity of each graph as well as the number of communities for each snapshot. Regardless of the number of communities, modularity reaches high values, thus suggesting dense connections between the nodes within communities but sparse connections between nodes which belong to different communities.

In total, the number of communities with very few members outweighs the number of heavily populated ones (Figure 6), which makes sense since most people are circumstantial users. However, there are also persistent users who appear on almost every snapshot, thus leading to the creation of persistent communities. The focus of this framework is to discover these communities and extract newsworthy and event-related information from them. To this end, we define two temporal measures: **persistence**, as the characteristic of a dynamic community to make an appearance in as many timeslots as possible (i.e. overall appearances / total number of timeslots), and **stability** as the ability to appear in as many consecutive timeslots as possible disregarding the total number of appearances (i.e. overall consecutive appearances / total number of timeslots). We expect consistent dynamic communities to be both persistent and stable.

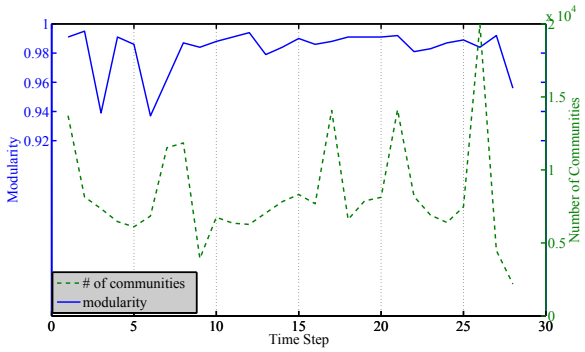

Figure 5: Modularity and number of communities for each timeslot.

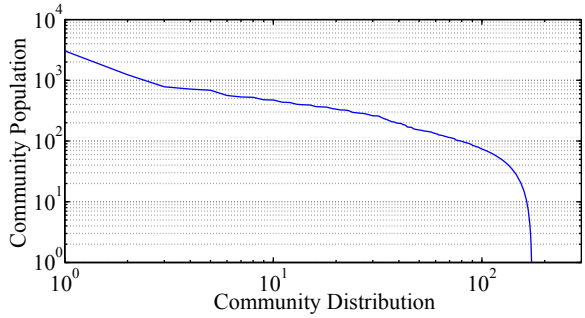

Figure 6: Distribution of the number of communities with respect to their population. Communities of trivial sizes (< 3) were removed.

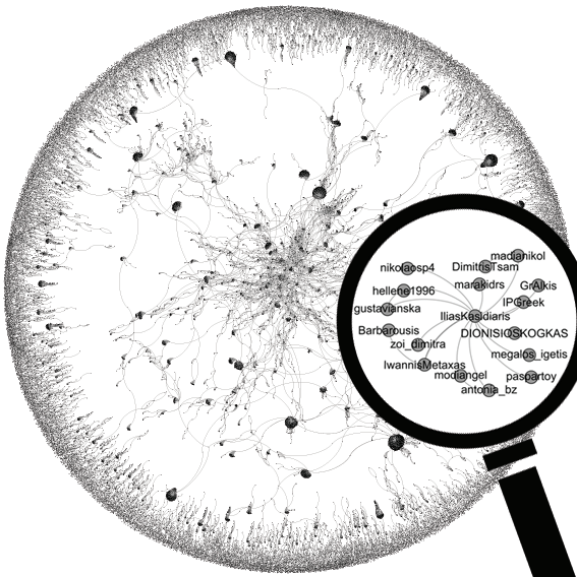

Figure 7: First timeslot interaction graph: The magnified section shows a Greek community commenting on a poll that presented the GGD party as the most popular amongst unemployed citizens. (Visualization created using the Gephi software [2])

The main challenge in this analysis pertains to the magnitude of data to be analyzed. The total number of extracted communities from all 28 snapshots is approximately 232K. Even by removing all the self-loop users and all communities with a population less than three, the amount of communities is reduced to 89K. In order for the reader to grasp the magnitude of available information, Figure 7 presents the graph produced from the first snapshot alone. Although there appear to be many distinct communities, the volume of data is too much for one to handle manually, especially when the desired communities are much smaller than the dominant ones. For instance, the magnified image in Figure 7 displays such a small community. Despite its size, it could provide an analyst with significant information: in this particular case a popularity poll has surfaced that ranks all Greek political parties amongst unemployed citizens. It is noteworthy that the central user (*@iliaskasidiaris*) as well as *@Barbarousis* are members of the Greek parliament and very active members of the GGD party.

Another example is displayed in Figure 8, in which a number of interconnected Greek and foreign communities are exchanging news and ideas. The graph contains active and former members of the Greek parliament (*@AdonisGeorgiadis, @thanosplevris*), local newspaper accounts (*@iefimerida, @enthemata*) and even a connection to the YouTube account; but through this graph a journalist or analyst could discover several influential users and bloggers *@neosklavos, @teacherdude* as well as find out important information from the commissioner of the human rights organization *@CommissionerHR*. Moreover, there are also several anti-GGD and anti-neonazi accounts (*@anti_xryshavgh, @AgainstNeonazi*) so the messages being dispersed vary from strictly informative to extremely sarcastic. Such groups exist all over the network and use many different languages, providing proof that the discussion and news about the GGD movement attracts the interest of people of different nationalities.

The reason for the great magnitude divergence between the small elusive communities and the dominant ones, is that the people discussing the topic in Greece are extremely fewer compared to the rest of the world. Moreover, the words Nazi or far/extreme right in English are popular and are used in various other contexts. For example the term "Grammar Nazi" is widely used to describe a person who is strict about grammar usage in a variety of memes. This creates an abundance of false positive extractions of such messages that could be avoided by introducing term exclusions to the crawler. However, such a filter would require human intervention and thus remove part of the system's implicitness and automaticity. Hence, the false positives along with the crushing difference in language usage popularity make the chances of discovering small communities even fainter.

An indication of this divergence is presented in Table 2 where the distribution of the 20 most popular languages of all the messages is displayed. English is the most dominant language, followed by Spanish. The rest of the languages exhibit a far smaller frequency. Although a language-based filter could provide a strict analysis of the communities per language, it would also result in the loss of a large number of posts, users and communities that could potentially link local to global communities and thus have a significant role in the overall study. In fact, it is of great interest to see how these Greek communities evolve and how they are connected to respective global communities. As an example, in the first snapshot there are distinct Greek communities, including members of the Greek parliament, which are con-

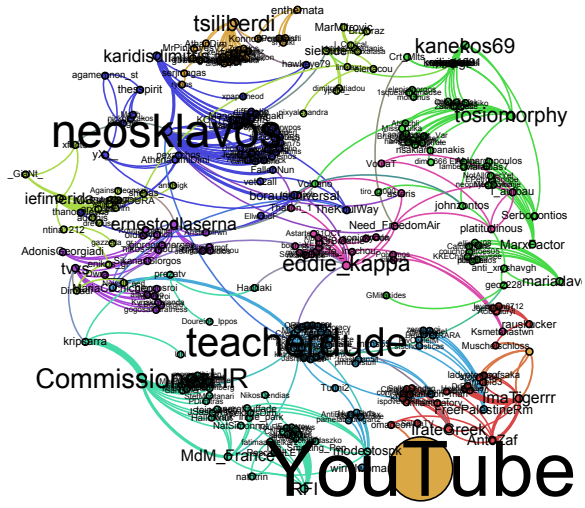

Figure 8: Example of a group of interconnected foreign and Greek communities surrounded by an abundance of groups and single users. (Visualization created using the Gephi software [2])

Table 2: Distribution of the 20 most popular languages from all the messages in the dataset.

EN	ES	IT	PT	DE	FR
556261	178211	50717	45516	36477	26631
ID	**NL**	**TR**	**HR**	**EL**	**ET**
23916	20093	16651	11868	11553	8925
SW	**AF**	**SL**	**SV**	**NO**	**SO**
8687	8611	8337	6586	6507	6191

nected through different users to a Greek, a German and a British newspaper account, to a war history account and to a number of politicians' accounts. Interestingly enough, the interaction between them in most cases is bidirectional which means that news and events are exchanged and dispersed by both sides.

An additional deduction from studying the snapshots is that in smaller communities, users' high activity does not necessarily imply a community with a high out-degree. Hence, the question remains about which communities, and as an extension, which users to follow to extract valuable information from a network concerning a specific topic. The answer lies in the evolution of these communities. Useful information can be extracted by studying the behavior of these communities (i.e. persistence, stability). This specific dataset presents a total amount of 89K communities of which 82K remain unevolved; this results in 7K instances of community evolution that are presented by the heatmap of Figure 9.

Overall there are 1.1K unique evolving communities, from which the most populated sustains a population of approximately 400 people $T_{6,16}$ and the most persistent $T_{312,3}$ spans over a period of 10 timeslots. The context of these two extremes could be of interest to any analyst or reporter. In the first case ($T_{6,16}$), there are two separate events that triggered this massive evolution. The first regards the fact that a person named Adolf Hitler is running for the elections in India and the second involves the statements of an ex-gay advocate saying that pro-gay school groups are like the KKK. The

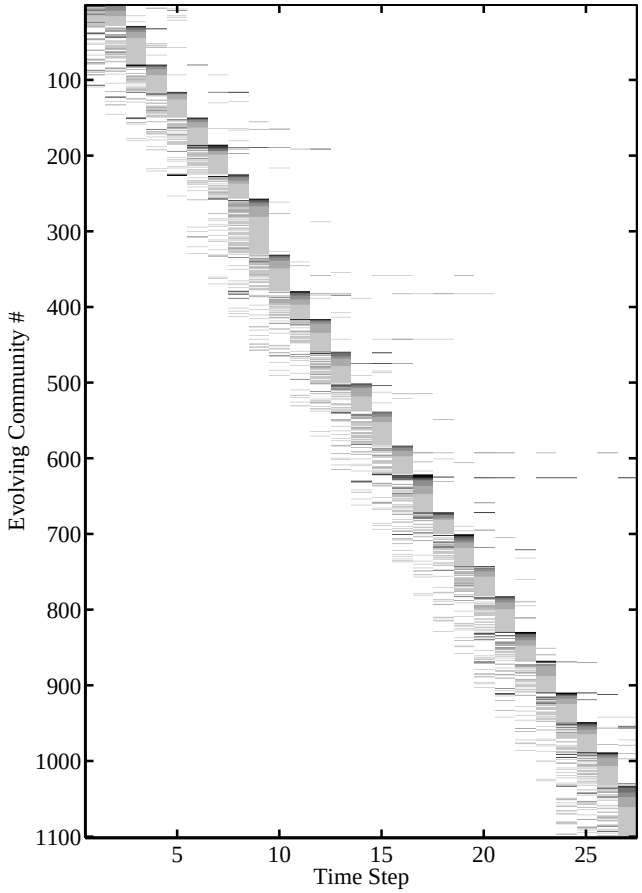

Figure 9: Heat map of each unique evolving community for each timeslot. The shades of grey indicate the size of the specific community. The population increases as the shade darkens.

interesting part is that all 798 mentioning posts were probably triggered by a tweet sent out by two different, though of similar context accounts (*YourAnonNews, AnonymousFM*) and received mentions by numerous different users in a period of two days. In the second case ($T_{312,3}$), the evolution of this relatively small community (maximum population of 11 people) spans timeslots 3 to 12, translating to the interval from the 6th to the 17th of February. The discussion in hand was in Spanish and involved a book presenting Hitler as an alchemist of emotions.

With respect to the Greek communities, we discovered many fairly small communities, scattered in time that are associated with a variety of events and opinions around the GGD party. One of the extracted discussions using the evolution analysis from dynamic community $T_{1288,12}$, concerned the wages of the members of the Greek parliament and how the reaction from the GGD party managed to introduce cuts. This dynamic community comprises only three members but due to its evolving state it stands out from the crowd of other minute communities and provides us with this newsworthy piece of information.

The above were only few examples of the news-related communities springing out of the perplexed Twitter canvas. By using the framework presented in this paper, a lot more can be revealed. However, we also need to stress the com-

plexities arising when analyzing the data. A first challenging problem was to identify and extract only the relevant communities about a specific topic: in our case, we found it hard to separate GGD-related communities from the rest, without using the tweet language or account information as a criterion. A further challenge pertains to the summarization of community information: currently, we are limited to an interactive analysis, i.e. exploring the graph and zooming in the areas of the graph that look interesting. In the future, we plan to develop and test automatic community ranking and selection approaches to speed up the analysis process.

5. CONCLUSIONS

This paper presented a framework for the effective analysis of the community structure, interaction and evolution in OSNs. In addition, an experimental analysis was performed on an evolving network extracted from user interactions (mentions) in Twitter. When applied on this network, our method uncovered a large number of dynamic communities with various evolutionary characteristics. The conducted experiments highlighted the potential of the proposed framework for discovering newsworthy pieces of information and real-world incidents around topics of interest. They also revealed the complexity of the analysis process due to the large magnitude of the data to be analyzed.

Future work includes the utilization of additional community detection techniques, the application of the framework on even larger datasets and the development of methods for automatically selecting and browsing through the discovered communities based on a variety of different evolution metrics (such as persistence and stability).

6. ACKNOWLEDGMENTS

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

7. REFERENCES

[1] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 913–921, NY, USA, 2007. ACM.

[2] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, editors, *ICWSM*, pages 361–362. The AAAI Press, 2009.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008 (12pp), 2008.

[4] A. Burns and B. Eltham. Twitter free iran: an evaluation of twitter's role in public diplomacy and information operations in iran's 2009 election crisis. In F. Papandrea and M. Armstrong, editors, *Proceedings of Communications Policy & Research Forum*, pages 298–310. Network Insight Institute, University of Technology, Sydney, 2009.

[5] R. Cazabet, H. Takeda, M. Hamasaki, and F. Amblard. Using dynamic community detection to identify trends in user-generated content. *Social Netw. Analys. Mining*, 2(4):361–371, 2012.

[6] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, editors, *KDD*, pages 554–560. ACM, 2006.

[7] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[9] M. Giatsoglou and A. Vakali. Capturing Social Data Evolution Using Graph Clustering. *Internet Computing, IEEE*, 17(1):74–79, Jan. 2013.

[10] D. Greene, D. Doyle, and P. Cunningham. Tracking the evolution of communities in dynamic social networks. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '10, pages 176–183, Washington, DC, USA, 2010. IEEE Computer Society.

[11] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[12] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, and et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[13] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM.

[14] K. McKelvey, A. Rudnick, M. Conover, and F. Menczer. Visualizing Communication on Social Media: Making Big Data Accessible. In *Proc. CSCW '12 Workshop on Collective Intelligence as Community Discourse and Action*, pages 46–50, 2012.

[15] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446:664–667, 2007.

[16] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media - performance and application considerations. *Data Min. Knowl. Discov.*, 24(3):515–554, 2012.

[17] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5:1–34, 1948.

[18] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 717–726, New York, NY, USA, 2007. ACM.