A Multi-Resolution Approach to Characterize the Connectivity Structure and Evolution of Large Graphs

Soheil Jamshidi
Department of Computer and Information Sciences
University of Oregon
jamshidi@cs.uoregon.edu

June 25, 2016

Abstract— **Graphs are widely used to represent the structure of large networked systems such as Online Social Networks (OSN). These graphs have a large number of evolving nodes (i.e., users) and edges (i.e., relationships). It is important to have practical methods to capture and characterize the connectivity structure and evolution patterns of such networks to gain insights about the corresponding system. However, existing techniques for graph analysis have limited scalability, offer limited insight about graph structure, and often do not capture the graph evolution. In this study, we present a new multi-resolution method to characterize the connectivity features and evolution of large graphs. The main idea is to divide the graph into a manageable number of meaningful elements and characterize inter and intra-element connectivity. We focus on the subgraphs of high degree nodes, i.e., core nodes, and identify the community of core nodes, i.e., core communities, as the main elements of the graph. This method allows us to perform the following analysis: 1) spatial analysis to determine how the size of core subgraph affects the number and characteristics of core communities which control the resolution of our analysis, 2) temporal analysis to characterize the evolution patterns of the core communities over time, and 3) spatiotemporal analysis to relate the spatial and temporal analysis. We use 14 snapshots of Google+ OSN to illustrate the capabilities of our approach.**

I. INTRODUCTION

Graphs are used to represent the structure of interconnected systems within different domains such as genomic in biochemistry, scheduling in operations research, and online social networks in computer and social sciences. This allows researchers to characterize the underlying network structure by examining the connectivity features of the graph including degree distribution, PageRank [11] and shortest path among nodes. While there are many techniques in complex network analysis to extract characteristics of graphs, they provide either aggregated graph-level measures (e.g., average node degree, diameter, and average path length) or node/edge-level attributes (e.g., degree, clustering coefficients, and assortativity) that do not provide sufficient insights about the graph. More specifically, two graphs might have similar aggregate features while their connectivity features are very different. Moreover, in many cases the system structure evolves with time and it is important to capture and characterize the evolution of its structure over time. The increasing presence of large graphs, with hundreds of millions of nodes and billions of edges in various domains such as OSNs, makes the mentioned problems more apparent.

One approach to extract the connectivity features of a graph is to focus on a coarser view of the graph. Instead of analyzing hundreds of millions of nodes, by using Community Detection Methods (CDM), a set of tightly connected nodes can be grouped as a community. In this schema, individual nodes can be replaced by communities and structure can be analyzed at the community level. The smaller number of elements make the analysis more feasible and manageable. However, there are some issues with this approach. First, CDMs cannot scale well to the size of large graphs. In order to use them, we need to summarize the graph with the cost of losing part of the information. Second, CDMs are non deterministic in most cases. Having the same set of data, the output might be different in different runs. Third, assuming that a CDM can run properly on the input graph, there will be hundreds of communities in a graph with tens of millions of nodes (as we illustrate in Fig. 6), which is a large number of elements for analysis, specifically for conducting temporal analysis of a graph. In this study we present a new approach to characterize the structure of large graphs as well as their evolution over time. The key idea is to divide the graph into a manageable number of meaningful units by focusing on high degree nodes (i.e., core subgraph) and find coarser views of these core subgraphs and consider them as our analysis units. Then, we characterize inter and intra-connectivity of each unit. Focusing on high degree nodes, considering that they are more central and play as connectivity hubs in the graph's structure, is a key step to overcome the described issues with current approaches. These core communities are meaningful and important units for characterizing a large graph as each one represents a tightly connected community of important nodes. More specifically, we can view a large graph as a collection of core communities along with their low degree friends and followers.

We use 14 snapshots of Google+ connectivity structure to examine and illustrate the capabilities of our approach. We focus on the nodes with the highest number of followers (in-degree) the we refer to as *core nodes*. Then, we use Combo [13] community detection method that relies on multi-objective optimization to identify communities within this subgraph. The output is the communities of core nodes that we also refer to as *core communities*. Most of CDMs are non-deterministic that results in community mapping variation. To minimize this effect, we run the community detection technique on the core subgraph n times. Then, we compare the communities that each node were mapped to (a vector with n values) and identify groups of nodes that have identical mapping vectors. We refer to each group of core nodes as *resilient communities*. The main tuning parameter for this approach is the number of high degree nodes, i.e., size of the core subgraph. This could possibly change the number of communities, and therefore change the resolution of our view. We refer to each core subgraph as a *view* and start from the top five thousand most followed nodes. We then we incrementally double the view size up to 40 thousands nodes. Having multiple views, similar to what is illustrated in Fig. 1, helps us to study the effect of adding more core nodes (spatial analysis). Furthermore, for any given view, we examine how connectivity features, mainly the core communities, evolve over time as the identity of the nodes and edges in the core subgraph changes (temporal analysis). These analyses provide complementary patterns of changes for individual communities. Therefore, adding these patterns together allows us to capture richer and more comprehensive patterns (Spatiotemporal analysis).

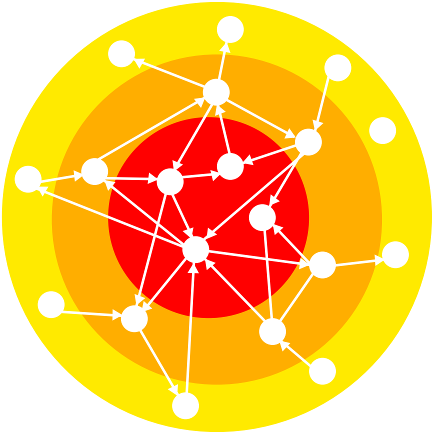


Fig. 1: Multiple views (core components) of graph based on number of followers- nodes in the red zone have more followers.

Having multiple views and applying spatial analysis, we are able to identify "how individual communities change as more nodes are added to the core subgraph"; Whether they split or stay the same or merge with other communities. By considering communities in each view over time, we can study "how communities in a single view evolve as the network changes over time"; whether they stay unchanged or their nodes churn a lot. Using the Spatiotemporal analysis we can observe "how applying different combinations of spatial and temporal mappings could result in different mapped communities".

The rest of this study is organized as follows: First, in section II we give an overview of the graph evolution analysis problem. we then discuss the earlier efforts in this domain. In section III we describe our datasets and discuss their basic characteristics. Next, in section IV we explain our methodology and how it would be beneficial to analyze large scale graphs. After characterizing the core analysis elements in section V, we define three angles of our analysis which consists of Spatial (section VI), Temporal (section VII), and Spatiotemporal analysis (section VIII) and their related outcomes.

II. Related work

Graph evolution is studied in a wide variety of networked systems such as social networks, WWW, and biological networks. Such networks are evolving over time by addition or removal of entities (nodes) and interactions among them (edges), which results in structural changes in the network over time.

There are different approaches to study the network evolution, such as maintaining the current status of a network and studying the patterns of network evolution. While updating the structural groupings and related metrics such as PageRank [11], clustering coefficient, and average degree help us to have a fresh view on the network at any time, capturing the evolution patterns such as merging or splitting groups of a network over time can help predict the future changes. In this section we review both categories.

A. Characterizing the evolving networks

One of the first efforts in this domain is done by Leskovec et al. [10]. They consider 12 different datasets such as two citation graph for U.S. patents and email communication network. They mostly focus on aggregate measures and observe that networks get denser over time. As a result of this network *densification*, the diameter may gradually decrease, despite the popular expectation for it to increase. Furthermore, they demonstrated that a giant connected component emerges that covers almost all of the nodes.

In another study, Leskovec et al. [9] quantify the bias of new edges towards the degree and age of nodes with which they form relation. They report the results of a microscopic analysis on four large social networks, namely Flicker, Delicious, Yahoo! answers and LinkedIn with up to tens of millions of nodes and edges. They show that the fraction of edges (friendships) that are initiated by new users is very high. However, the number of interactions becomes uniform over time. Using various models such as preferential attachment along with their observations, they proposed a network generation model that captures evolution characteristics of real networks. Although these types of observations provide useful insights for the research on the graph evolution, considering the different types of networks and their scale we still need more practical analysis methods to capture more detailed evolution patterns.

There are other studies that study coarse views of the graph. Backstrom et al. [2] study the effect of structural factors on group (community) evolution and changes within a network. They analyze the structural factors that influence a user joining a community and expanding a community over time. They also study communities to realize whether users loose their interest in a group or whether a group label convinces them to join a new community. Their analysis on co-authorship network of DBLP and LiveJournal social network depicts that both the number of friends in a new group and their connectivity are important factors in attracting individuals to join a new community. They also report that structural factors could also be used to predict the way a community grows over time. Considering different types of communities, types of node behaviors and summarizing them while having hundreds of communities would be a difficult task. Furthermore, there might be other causes in different levels that affect the node membership in a community such as network level events.

Zhao et al. [16] analyze the dynamics of Renren social network over time at different scales including node, community and network levels. At the node level, they show that the preferential attachment model gradually weakens in modeling the edge creation as the network grows and matures. They show that the edge creation becomes increasingly driven by connections between existing nodes as the network matures instead of being related to the degree of edge destinations. At the community level, Louvain community detection algorithm [3] is used to track communities across snapshots. They observe that community mergers can be predicted with reasonable accuracy using structural features and dynamic metrics such as acceleration in community size. When a merger happens at the network level, it is observed that its impact is significant in the short term, but quickly fades with the constant arrival of new nodes. They also conclude that node level behaviors are not only driven by dynamic events at the node-level, but also are influenced by events at the community and network levels.

B. Capturing network evolution patterns

Although studying the evolution of aggregate factors of a network is important, for an in-depth insight, we need to analyze the evolution of communities as an element of network structure.

There are studies that focus on communities when analyzing the network evolution. Their main components are as follows: First, clustering method which is one of the community detection algorithms; second, similarity measure to identify a community in the next snapshot of the network; and finally, type of events that are identified based on the defined similarity measure. These components are also the main differences among them.

Asur et al. [1] propose a method for identification of critical events that occur in evolving interaction networks. Events are extracted by comparing the clusters which are detected using a clustering method on two consecutive snapshots of network in time T_i and T_{i+1}. Events include *continuing*, *splitting*, *merging*, *dissolving* and *forming* for communities. Furthermore, to analyze the influence of individuals' behavior on communities, they introduce *appear*, *disappear*, *join* and *leave* to cover the churn of individuals over time. Based on these events, they propose related measures for sociability, stability, influence and popularity which help rank the users in terms of link prediction and influence maximization. Although they indicate that their framework is independent from the clustering method, it is acknowledged that the optimality of the clusters will play a key role in the efficacy of obtained results.

Palla et al. [12] study the pairwise mapping of overlapping communities. In Palla's study, clique percolation method (CPM) was used to allow groups to overlap. Networks at two consecutive time frames T_i and T_{i+1} are merged into a single graph Q(T_i, T_{i+1}) and in each time frame groups are extracted using the CPM method. Afterwards, the communities in T_i and T_{i+1} match if they are in the same group within Q(T_i, T_{i+1}) graph. Matching is then performed based on the value of their relative overlap sorted in descending order. Possible events between groups are similar to previous studies.

Limitations such as using a specific type of clustering method, computational costs and coverage of all possible events with high accuracy are still open problems and many efforts have been made to solve it. Brodka et al. [4] proposed the Group Evolution Discovery method, *GED*, which uses *inclusion* similarity measure in order to identify what happens within a group in successive snapshots of a network. Despite *Jaccard index* which is used in similar study done by Greene et al. [7], *inclusion* covers both the quantity (the number of members) and quality (the importance of members) of the group which contribute to an event. The GED method was also designed to be more flexible compared to methods mentioned above and more accurate in finding all of the events and to fit into both overlapping and non-overlapping groups. Still, the final results are affected by accuracy of the clustering method and number of analysis units.

In a different study on Google+ dataset by Gong et al. [5], the interplay between user attributes and links is explored. The basic model is the union of a social graph with a bipartite graph capturing user-attribute associations. They extend the usual macroscopic network metrics (degree distribution, density, and diameter) to attribute-labeled graphs, and make several interesting observations about the impact of attributes on network evolution. Next, they define a model where node attachments are driven both by social connectivity and by attribute proximity. This model is shown to match the main macroscopic features of the data, and to perform slightly better than the closest generative model by Zheleva [17].

Among many efforts on snapshot-based analysis of networks, there are studies that track the evolution in an incremental way. Lee et al. [8] do not analyze the network snapshot by snapshot over time. Instead, they summarize the network into a skeletal graph based on density parameters and then analyze the stream of updates for this skeletal graph using a sliding time window. However, finding the thresholds that result in optimal outputs for different networks could be a complex task.

Recalculating the structural measures as a network evolves is a time-consuming task. Analyzing huge numbers of communities, mapping them and comparing them is still far from an optimal solution and is hard to handle. Furthermore, none of the reviewed studies propose a way to identify the communities in a meaningful way that enable us to map detected communities to social groups in the real world. Summarizing a network at different levels and analyzing it in different scales would be an efficient way to capture the most important changes of a network over time.

III. Datasets & Basic Characteristics

In this section we present our datasets and some of their basic characteristics.

A. Datasets

We have access to 14 snapshots of Google+'s structure that are crawled roughly one month apart, starting from August 2012 to May 2013 [6]. Each snapshot has a directed edge view in the form of $E = \{(v, w), v\ follows\ w\}$.

Fig.2 shows the number of nodes per snapshot, with the x-axis presenting the time of each snapshot, and each bar indicating the total number of nodes per snapshot as well as number of new nodes (green bar) and removed ones (red line). The network size ranges from 60 million nodes in the first snapshot to around 160 million in the last one. We refer to the nodes that are not in snapshot(i) but are part of snapshot(i+1) as *new nodes*. Nodes that are in snapshot(i-1) but are not in snapshot(i) are called as *removed nodes*. The fraction of new nodes in each snapshot is less than 15 percent. Considering that the rate of node addition is higher than node removal, network size is steadily increasing.

Fig.3 shows the number of edges per snapshot, with x-axis presenting the time of each snapshot, and each bar indicates the total number of edges per snapshot as well as the number of new (green bar) and removed edges (red line). The definition of new and removed edges is similar to the new and removed nodes. The number of edges varies in the range of 800M edges in the first snapshot up to 2.6 billion edges in the last one. The rate of edge addition and removal decreases over time, with at most 70 percent new edges in the fifth snapshot and 70 percent removed edges in the forth snapshot.

Despite the changes, a large fraction of overlap between nodes and edges in consecutive snapshots suggests that the network structure is stable.

Fig.4 plots the density of the graph per snapshot. Average degree is the overall effect of the number of edges divided by the number of nodes. In all of the snapshots, average degree fluctuates between 30 and 40.

B. Basic Characterization of Snapshots

We present some of the basic characteristics of network snapshots to illustrate that despite significant changes in the graph, some of its average features or distribution of features may not exhibit any measurable changes. The main problem with aggregate measures like average degree is that they can be easily misled by skewed data. For example, large numbers of low degree nodes can reduce the average degree even

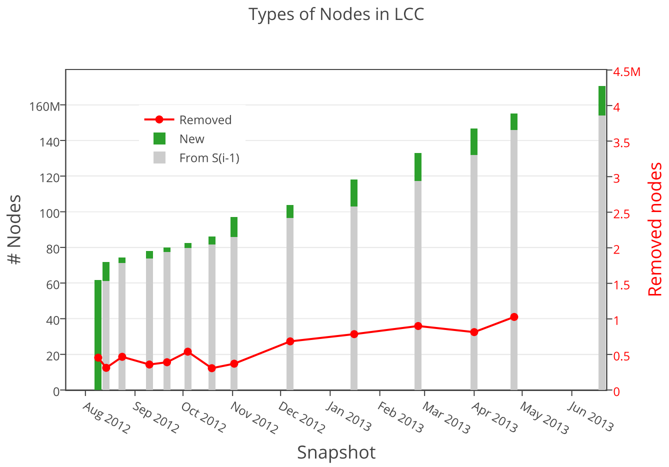


Fig. 2: Type of nodes, new, removed and from previous snapshot

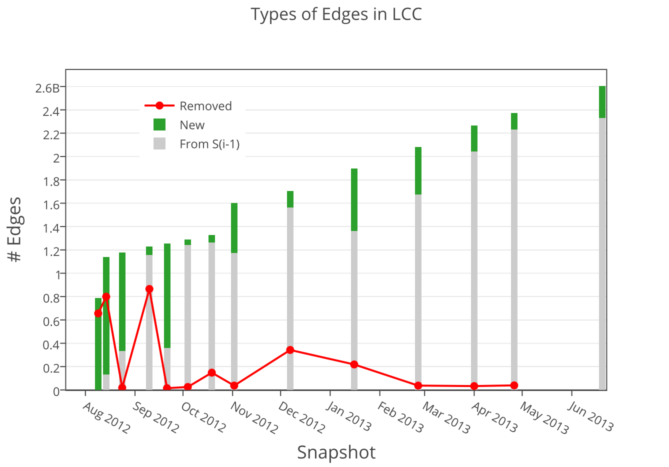


Fig. 3: Type of edges, new, removed and from previous snapshot

though there might be some large degree nodes in the graph that are important for us.

Fig.5-a and Fig.5-b illustrate the Complementary Cumulative Distribution Function (CCDF) of number of friends and followers per snapshot, respectively. The maximum number of friends goes up to around 10,000 nodes and around a million for the number of followers. Both distributions are skewed, i.e., Google+ has a small fraction of high degree nodes, with 70% of nodes having less than 10 friends and around 80% of nodes have less than 10 followers. Both distributions do not change significantly. Therefore, they do not reflect changes in the graph at a micro level.

Fig. 6 shows the number of communities that are identified using Luvain community detection technique [3] across all snapshots. The number of communities is very large, more than 50 thousands across all snapshots, and keeps increasing with time with a maximum of 340 thousands communities

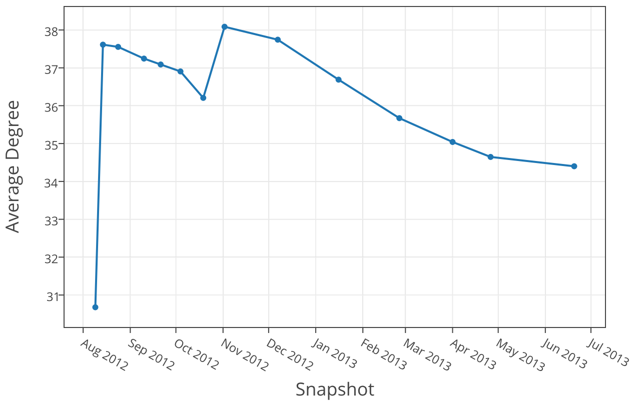


Fig. 4: Average degree for the graph

in snapshot 13. Given a large number of analysis units, community level analysis is still very complex.

IV. Methodology

The goal of this study is to characterize the connectivity features of large graphs. These features not only reveal valuable insight into the structure of a graph but they can also be used to assess the evolution of a graph's structure.

Towards this aim, the main idea is to divide the graph into a manageable number of meaningful units and characterize inter and intra connectivity of each unit. One possible approach to find a proper unit is to detect all of the communities within the graph and use them as a unit for our analysis. However, there are several issues with this approach. The size of communities typically is between 50 to 100 on average [10], thus a huge graph would have tens of thousands of connected communities which is still a large number for proper analysis of each unit. Furthermore, a significant majority of nodes in most of the communities are low degree nodes that are not important, i.e., they do not play an important role in the connectivity of the graph. For example, by a small number of connections to the high degree nodes, they often are hanging from the rest of the graph.

To overcome such issues, our key idea is to focus on high degree nodes since they are more central and play as connectivity hubs in the graph's structure. In the context of directed Online Social Networks (OSN), users with the most number of followers are clearly more important than nodes with more friends. Considering the fact that the degree distribution is very skewed in most graphs, this implies that a very small fraction of nodes have a large number of friends and followers. We refer to the nodes with the highest number of followers (in-degree) as *core nodes*. We consider the subgraph of core nodes and identify communities within this subgraph. These are communities of core nodes that we also refer to as *core communities*. This subgraph is not necessarily connected. For example, a high degree node that is only connected to a large number of low degree nodes, is not connected to the rest of the core nodes.

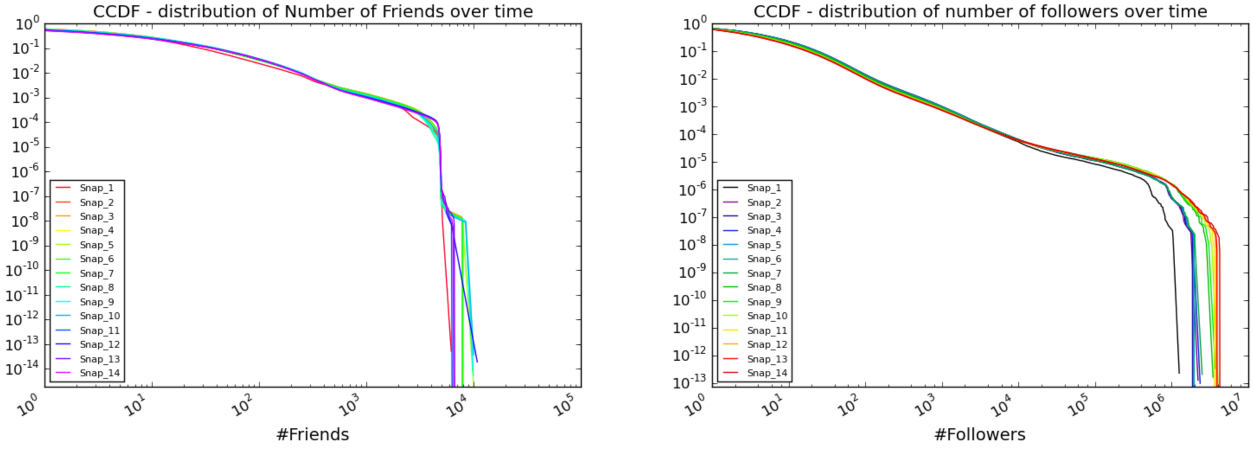
These core communities are meaningful and important units for characterizing a large graph as each one represents a tightly connected community of important nodes. More specifically, we can view a large graph as a collection of core communities along with their low degree friends and followers. Furthermore, we collect social and geographic attributes of higher degree nodes in each individual community to assess whether they exhibit any social coherency or similarity as well. This view allows us to characterize the connectivity in two tiers:

1) Intra and inter-connectivity within individual core communities.
2) Among core communities and their friends and followers.

The main tuning parameter for this approach is the number of high degree nodes, i.e., size of the core that we consider. This could possibly change the number of communities and thus the resolution of our view.

Altogether, our methodology consists of the following steps:

1) ***Identifying the core subgraph***: We select the top N nodes with the highest in-degree (number of followers). Since some of these nodes are not strongly connected, we focus on the Largest Strongly Connected Component (LSCC) of these core nodes as a directed graph. We refer to each top N subgraph as a *view* and we start from the top five thousand highly followed nodes, then we incrementally double the view size up to 40 thousands nodes. Having multiple views, just like what is shown in Fig.1, is an efficient way to study OSNs [15] and helps us to study the effect of adding more core nodes.
2) ***Detecting core communities***: We leverage Community Detection Methods (CDM) to identify core communities. The selection of seeds could change the output of community detection techniques such as Luvain [3]. To ensure that the detected core communities are meaningful, rather than a side effect of specific random seeds, we run the Combo CDM[13] on the core subgraph multiple times and identify groups of nodes that are mapped to the same community in all runs. We refer to these groups as *resilient communities*. Using SocialBaker [14], we collect social and geographic attributes of sample users in resilient communities. If a majority of nodes in a core community exhibit similar attributes, this indicates that the core community is indeed a socially meaningful unit in the corresponding OSN.
3) ***Extracting features of a snapshot***: The key features of each graph are the connectivity within and among core communities as well as their friends and followers.
4) ***Spatial analysis***: Since core communities are the key elements of the graph structure in our approach, it is important to track their evolution properly. This is a difficult task because the size and characteristics of core communities could be significantly affected



(a) Number of friends

(b) Number of followers

Fig. 5: CCDF of number of friends and followers in the entire graph

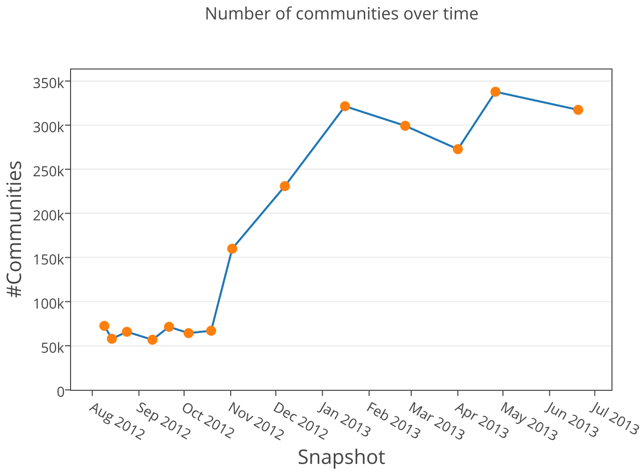


Fig. 6: The number of communities over time

by the size of the core subgraph. In particular, as we expand the size of our core subgraph, the newly added core nodes may connect to a single existing core community and may split that community into multiple communities, or connect to different communities and may cause them to merge, or they may form a new community. Therefore, in order to track the changes we expand the size of the core subgraph (view), incrementally and examine the above changes in the core communities as we change the core view. To conduct this analysis, we need to map an existing community C in a particular view V to a community C' in a larger view V'. Toward this end, we examine what fraction of nodes in C are located in each core community in view V' and use the fraction as a confidence for our mapping. We map C to a community C' in V' that contains the largest fraction of C's nodes and thus has the highest mapping confidence. Intuitively, multiple communities in view V can be mapped to a single community C in view V'. Discovered dynamics in community formation and evolution as we change the core size offers valuable insights about the connectivity structure among core nodes, which is the main goal of our methodology. We use *Sanky* diagrams to visualize the pattern of formation, merging, and splitting among the core communities, which offers the key insight. We also examine how various connectivity metrics for each community evolves as we expand the view.

5) ***Temporal analysis***: For any given view, we examine how connectivity features, mainly core communities, evolve over time as the identity of the nodes and edges in the core subgraph changes. Since we rank nodes based on their in-degree, rank of nodes differs in different snapshots as they attract different numbers of followers over time. It results in addition and deletion of nodes in top N over time which is different from spatial analysis where we only observe addition of nodes in the larger views. Similar to spatial analysis, we need to map communities across different snapshots. Our strategy for mapping the communities is the same as what we explained in spatial analysis. To illustrate the evolution of individual communities in a given view over time, we use Sankey diagrams. We also examine how various connectivity metrics for each community evolve over time.
6) ***Spatiotemporal analysis***: Spatial and temporal analysis present orthogonal and complementary patterns of changes for individual communities. Therefore, adding these patterns allows us to capture richer and more informative spatiotemporal patterns. For example, given two dimensions of pattern evolution, one could look into how a core community in the smallest view of the first snapshot evolves after five snapshots and in a

larger view. The status of community C in view i of snapshot j can be reached in multiple ways as follows:

- **Temporal-then-spatial**: Using the temporal evolution of C, we can determine its mapping in snapshot j. Then, using the spatial pattern, we can derive the corresponding community in view i.
- **Spatial-then-temporal**: Using the spatial pattern, C's mapping for view i is determined. Then, using the temporal pattern of that community, we can derive its mapping in snapshot j.

There are certain number of core communities at view i of snapshot j. Therefore, the key question is whether different mapping strategies result in different mapping outcomes or not.

V. Core Communities

In this section, we focus on the core subgraph with different sizes, i.e., different views, and the characteristics of core communities.

Node and edge reachability, To demonstrate the importance of core nodes, we show the percentage of reachable nodes and edges from core nodes. We consider cores with size starting from 100 up to 300 thousands within snapshot 9. Fig.7 plots the results of node and edge reachability for different core sizes. It suggests that by using a view of 5000 high in-degree nodes (0.005 percent of all nodes in snapshot 9), around 20 percent of nodes and 45 percent of edges are reachable. Increasing the size of core nodes to 40K results in reaching 25 percent of nodes and 53 percent of edges. However, expanding the core would not necessarily improve reachability; as we can see if we expand the core size from 140K to 300K, we increase our reachability to 3(0.5) percent more nodes(edges) respectively. Considering the top N views, Fig.8 depicts the fraction of nodes and edges that are reachable from each view over time. These observations confirm that considering a small number of important nodes results in accessing a large fraction of the graph.

Core stability: Fig. 9 plots the Cumulative Distribution Function (CDF) of the number of snapshots that nodes appear among core nodes. This illustrates that a majority of nodes, 60 percent of them, were among the core nodes for more than half of the snapshots. 40 percent of core nodes were in the core in all 14 snapshots. This shows the stability of core nodes over time. Furthermore, the ratio of covered core nodes by all core communities is between 94 to 98 percent. This fraction generally grows with the size of the view.

Size of core communities: Fig. 10 shows the size distribution of resilient communities. Each dot represents a community and the y-axis shows its size. Size distribution clearly shows that nodes are not grouped among the communities evenly. There are a small number of them which have a large number of nodes, at most 40 to 30 percent of nodes as we expand the view and a large number of communities with at least a fraction of a percent of nodes. As we expand the view, number of small communities that cover less than 15 percent of nodes increases.

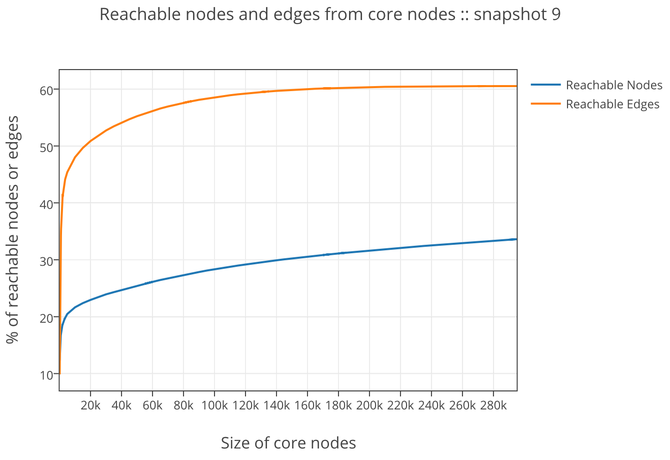


Fig. 7: Node and edge reachability - snapshot 9

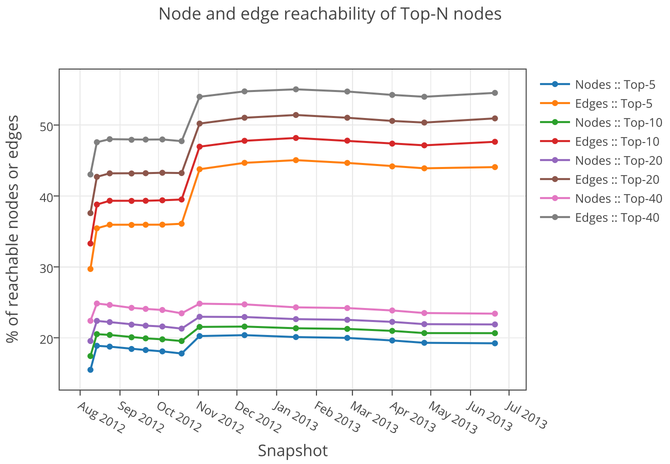


Fig. 8: Node and edge reachability of top nodes

Geo-Social context of core communities: To make sure that our communities are aligned with the social context of Google+, we crawled user attributes from SocialBakers, which provides three tags for each popular node, including a country code and two other tags that specify the types of users. Tags that are more popular among nodes of each community are selected as the community's social label. Among the crawled tags, three tags are popular: community, celebrity and entertainment. To choose more specific social labels, we ignore these three tags, and select the next most popular tag as the user type. For example, a community with ”US_Artist” label, implies that the majority of its nodes are ”artists” from the United States. We refer to each community using these labels in the rest of the paper. Fig.11 shows the number of nodes that are tagged with a user type and also their country codes for one of the core communities. After ”celebrity” tag, which is one of the general and popular ones, ”Artist” is the most popular tag for the user type (22 percent confidence) and the most popular country tag

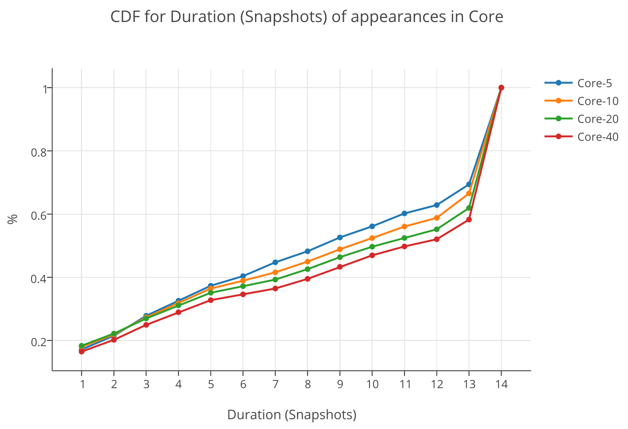


Fig. 9: CDF Duration(snapshots) of appearances in Core

is US. Therefore, this community is labeled as ”US_Singer”. Footprints of other communities are available in our project’s website[1]. In Table.I we report the core communities of the smallest view along with their size and the number of nodes that have social tags. We also report the fraction of nodes that their country code or their user type match with the community label as labeling confidence. For example, out of 30 labeled nodes of JP_Singer community, all 30 nodes have either ”JP” as their country code or ”Singer” as their type. The fraction of labeled nodes is not very high since SocialBaker has information for a limited number of popular nodes. However, considering the labeling confidence and also based on our manual profile checking procedure, we can make sure that the core communities are socially coherent.

TABLE I: Geo-social labels for core communities, their size and labeling confidence for Top 5K in the first snapshot.

Geo-Social label	Size	Nodes with tags	Labeling confidence (%)
US_Singer	459	128	32.03
TW_Club	175	56	33.93
JP_Singer	187	30	100.0
TH_Star	34	7	71.43
US_Artist	624	203	52.71
US_Writer	1097	521	60.84
ID_Life	21	7	71.43
VN_Life	82	14	78.57

Connectivity of Cores: We visualize four different types of communities in Fig. 12. The color of each node is related to its degree. High degree nodes are darker. Fig. 12-a shows how nodes of a community labeled as TW_Club are connected to each other. There is a tightly connected component in that community which is connected to two smaller components. It’s interesting to note that part of this community splits into another community as we add more low degree nodes. Fig. 12-b, JP_Singers, shows how these nodes form a super connected component and are tightly connected to all other nodes. These nodes always stay together and are reluctant to merging with or splitting into other communities as we expand the view over time. Fig. 12-c is one of the most central and largest communities with many dark nodes in the center. There is no doubt that there are many low degree nodes and communities that are interested to be connected to this community. And finally, Fig. 12-d shows how we were able to find a small number of Thai celebrities that are connected to each other and are loosely connected to some other high degree nodes.

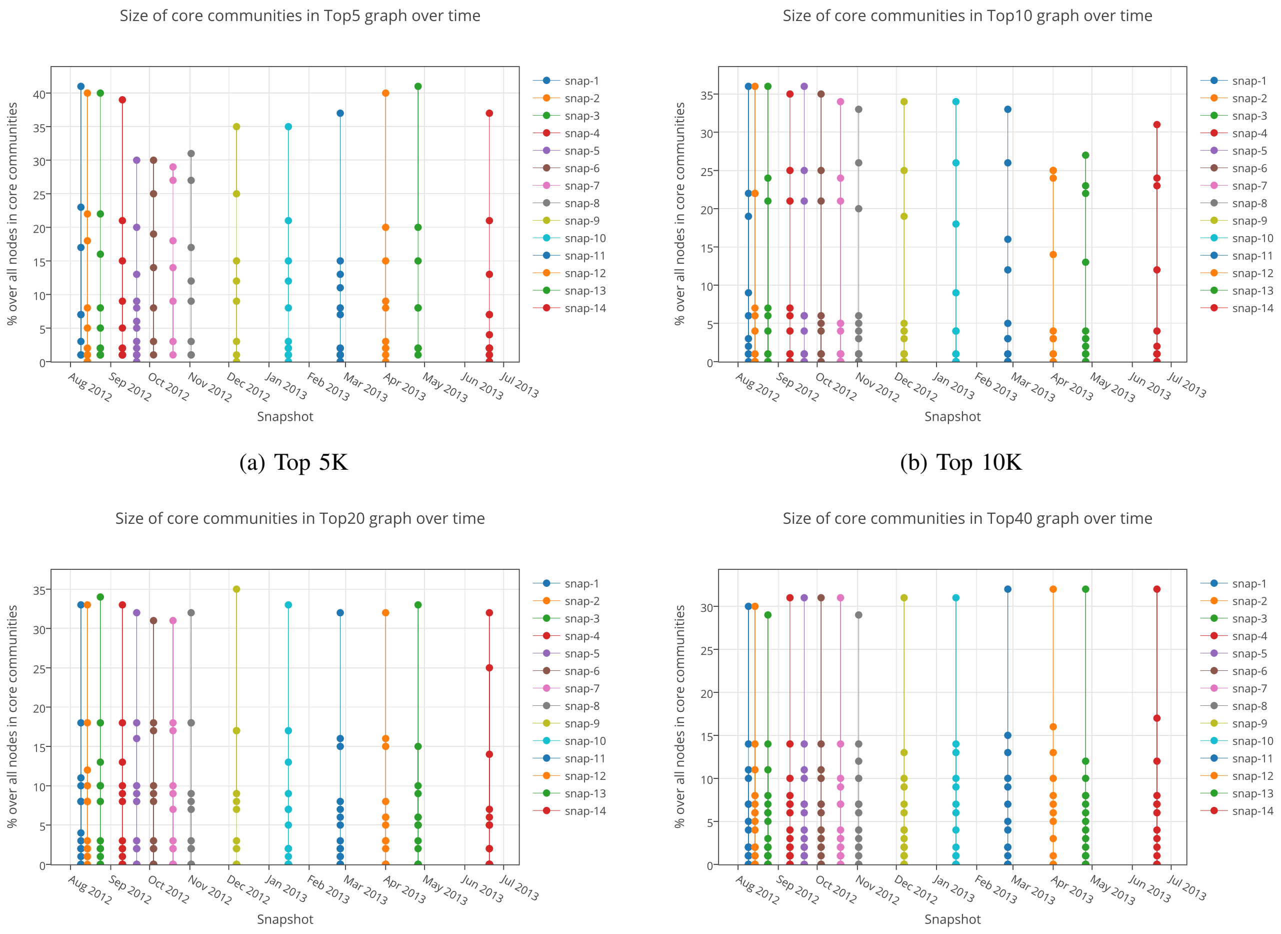
Inter-community connectivity is also another interesting aspect of our analysis. Fig. 13 shows how communities are connected to each other. The size of circles are relative to the size of communities, and edge thickness is relative to the number of edges that connect two communities together. For each community, we color the edge with the maximum weight as red. As it is depicted, most of external edges point towards the *US_Writer* community. The majority of nodes in all of the communities except *JP_Singers* are connected to the nodes of the *US_Writer* community. However, Japanese singers are more interested in US singers. *US_Artist* and *US_Writer* are tightly connected and our observation of other views shows that connections among communities increases as we add more nodes.
We grouped all of the nodes that are in top N graphs but are not part of resilient communities and labeled them as *unstable* nodes (yellow circle). Fig. 13 depicts that connections among this type of nodes and communities is not notable.

In the next three sections, we will analyze resilient communities over time and in different views to uncover their evolution characteristics and patterns.

VI. Spatial Analysis

In this section, we focus on changes in individual communities as we expand the view. Intuitively, node addition may result in expanding, splitting or merging of communities. We use Sankey diagrams in order to show the evolution as we expand the view. Sankey diagrams are a specific type of flow diagrams, in which the width of the arrows is shown proportionally to the flow quantity. Fig. 14 is the Sankey diagram that shows how communities evolve as we expand the view in snapshot 10 (Jan 2013). In this diagram, horizontally from left to right we expand the view (core size), so there are 4 columns for top 5, 10, 20, and 40K views, respectively from left to right. Each colored rectangle represents one of the core communities and the width of links between them shows the fraction of nodes that are mapped to a particular core community in the bigger view. We use a threshold to manage the minimum fraction of mapped nodes between two communities that should be displayed in this graph. By using a small threshold, we can track the small fractions of nodes that are mapped between any two pairs of communities as well. However, using a

[1] Our project website: http://onrg.cs.uoregon.edu/soja/Projects/MRA/



(a) Top 5K

(b) Top 10K

(c) Top 20K

(d) Top 40K

Fig. 10: Size distribution of resilient communities with $size > 9$

small threshold results in more links among entities that makes it hard to understand the patterns. As we expand the view, some interesting patterns emerge. For example, a couple of Media related communities that appear in the top 20 (Fig. 14-A), communities of Russians and Iranians emerges in the top 20 and continue in top 40 (B). There are also relations that emerge in different views. For example, relation among core community of JP_writers that split from core community of Taiwanese in top 10, merges again in top 20 and both merge to JP_Singers in top 40 (C, D, and E respectively). These are the examples which demonstrate that considering a single view can results in missing a part of knowledge. In the following subsection we categorize the core communities based on different types of spatial evolution.

A. *Spatial evolution patterns*

The core communities show different spatial evolution patterns. We observe the following main evolution categories:

- **Long Live Communities**: Some of the communities are identified in all of the spatial views. A fraction of these communities are also stable, meaning that they do not merge with or split into other groups. They have high mapping confidence as a result of stability of their core nodes. An example of such behavior is the core community of Japanese school girls, labeled as JP_Singer which covers around seven percent of view size.

This community has a strong two-way connection with the core community of Taiwanese. This group has the highest Clustering Coefficient in top 5 and top 10 and the second best in top 20, indicating that friends and followers have tight connections with each other as well. Its size (Fig. 10) and density stay the same as we add more nodes, which implies their low degree friends and followers have the same characteristics. Another fraction of these communities, such as a community of Vietnamese, labeled as VN_life, grows and gets denser as we add more low degree nodes in the bigger views but is reluctant to merge with others or split into smaller communities. The third group of these communities change in most aspects and also merge and split as we expand the view. Br_(Media—Singer) that is also

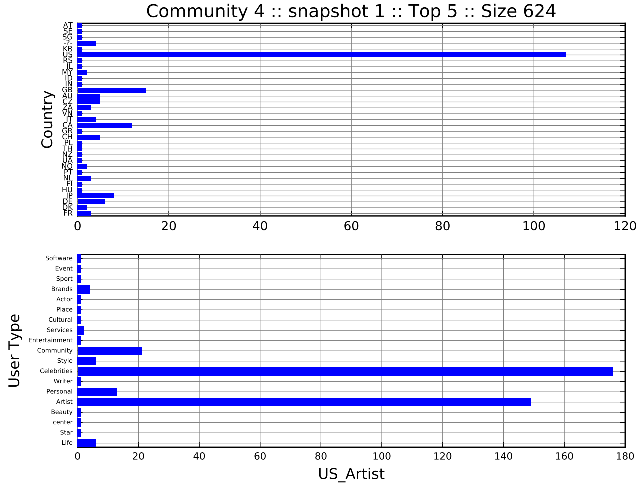
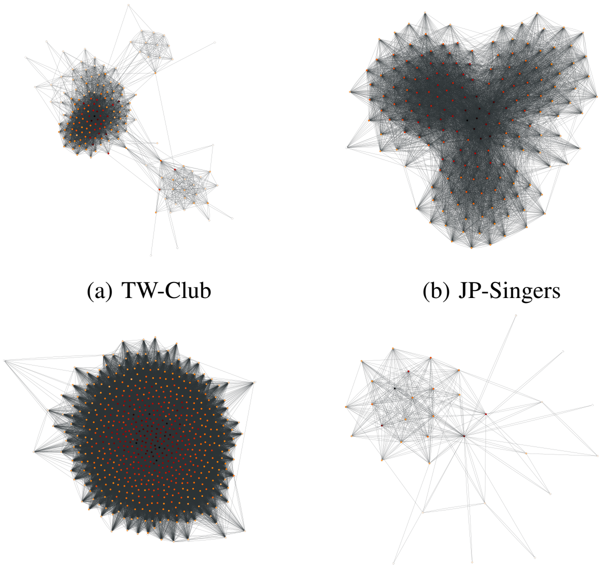


Fig. 11: The geo-social footprint of US_Artist core community



(a) TW-Club

(b) JP-Singers

(c) US-Artist

(d) TH-Stars

Fig. 12: Visualization of communities in the first snapshot

labeled as US_(Brand—singer) is the best representative for this group.

- **Some groups emerge and keep growing**. For example, the Russian community, Iranian community, and a core community that is labeled as BR_MEDIA. They start in top 20 and stay in top 40 as well. They increase in size and get denser as we expand the view.
- **Split from others**: Some split from other communities and merge with other communities a lot. Small media related communities such as Fr_Media and It_Media are the best examples for this kind of communities. On the other hand, some show interesting characteristics. AU_Sport community is very small in size with a small average degree compared to the rest of the communities but it appears and remains visible in the larger views. It has the highest Clustering Coefficient in top 20 and top 40 before it merges into other communities. It has the lowest churn rate among its nodes, meaning that nodes remain in the core graphs for a larger number of snapshots.
- **Some communities are more central**: US_(Artists and Writers) are the best examples for this category. They are the biggest communities and cover 25-40% of nodes in each view. Most of external edges connected to them have the highest Average PageRank across all views. Other communities tend to merge with them more than others and they are connected to a large fraction of the graph.

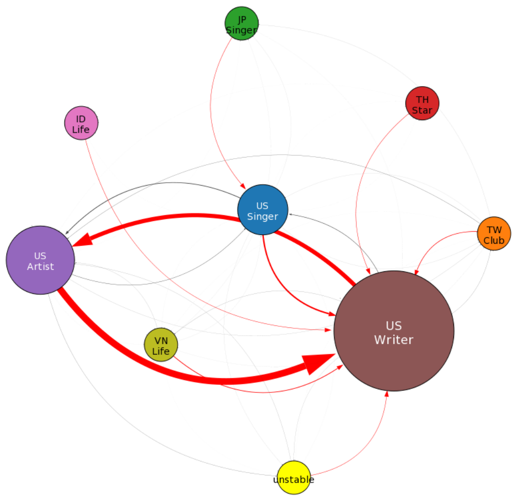


Fig. 13: External connections of communities in top 5 view - snapshot 1

VII. Temporal Analysis

In this section, we consider the evolution of communities over time. In addition to new nodes, some nodes are also removed due to changes in the number of in-degree or being removed from the graph. These changes result in merging or splitting or emergence of new communities. Fig. 16 shows the temporal evolution of communities in the top 10 view. In temporal analysis just like spatial, we can manage the level of details using a threshold. This threshold is the minimum fraction of nodes from a community that map to another one in the next consecutive snapshot. Temporal patterns that emerge over time can be categorized as follows:

- **Shrinking over time**: Some communities such as US_Artists, Fig. 16-A, are mapped with high confidence over time in all 14 snapshots without any major split or merge. However, they are shrinking in size, suggesting that some other high in-degree nodes replace a part of its members over time. Their

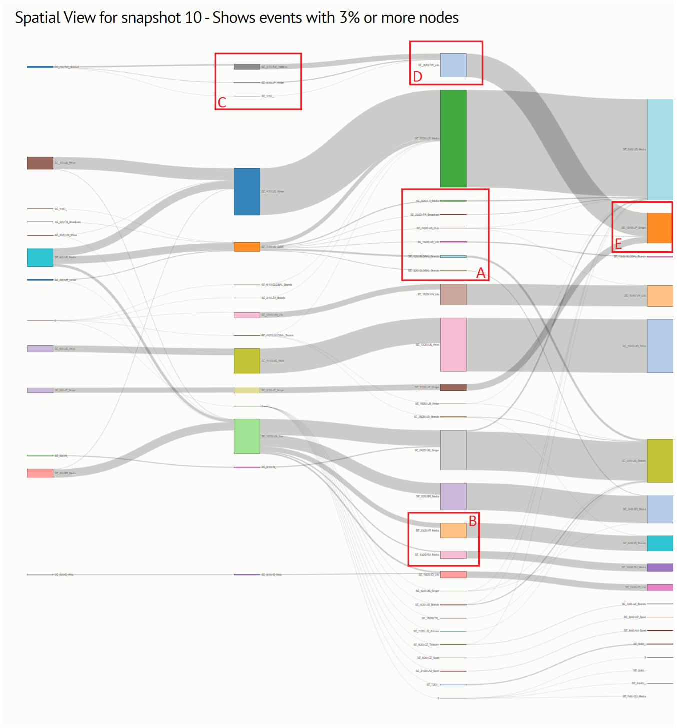
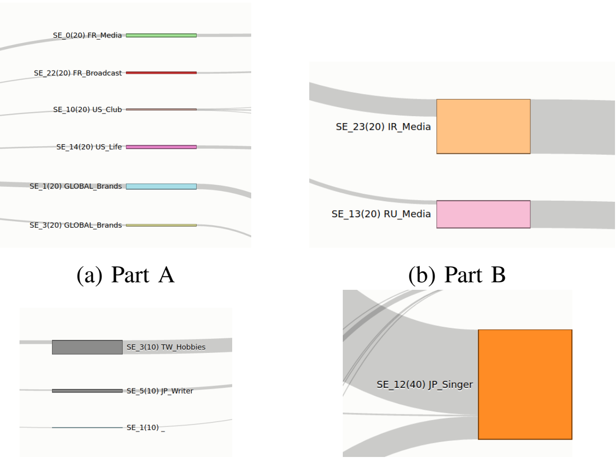

Spatial View for snapshot 10 - Shows events with 3% or more nodes
C
D
A
E
B

Fig. 14: Spatial view for snapshot 10, including all node mappings more than 3% of source nodes



(a) Part A

(b) Part B

(c) Part C

(d) Part E

Fig. 15: Magnified segments of spatial view (Fig.14)

clustering coefficient is increasing over time with a jump in snapshot 11 (March 2013), suggesting nodes are connecting more frequently with each other over time. They have the best average degree in all of the snapshots, indicating that they are the densest community in this view. Members of this group stay among core nodes frequently, they are the second best after JP_Singers though.

- **Combination of big communities**: US_Writer, Fig. 16-B, is the biggest community in this view over time. This group of nodes are split into other major communities such as US_Media and US_Singers. They are the second densest community over time with average degree around 120. Nodes are reluctant to triangle connections and their clustering coefficient is one of the lowest ones. However, they were among the core nodes for most of the snapshots.

- **Stable and isolated**: Fig. 16-C, labeled as JP_Singers, are a community that are tightly connected to each other, not interested in merging or splitting to other communities and stable in number of nodes (+- 20). The clustering coefficient of this group as well as their density is stable over time and they are the third densest community in this view. They have the most stable nodes in terms of appearance among core nodes, having 70 percent of nodes in the core for more than 12 snapshots out of 14.

- **Drops make the ocean**: In the first five snapshots, Fig. 16-E, a couple of small size communities merge and form the US_Media community. These merges also happen in the top 10 view in snapshot 10. Nodes and edges that are added in these snapshots, change formation and bring all of these communities together. US_Media community form the biggest community starting from snapshot 8 to the end.

We observe one of the interesting temporal patterns in snapshot 11 (March 2013). Part of three communities, split in this snapshot, and then merge again to their parent in the next snapshot (Fig. 16-D). A community of Japanese writers split from Taiwanese, French artists split from US artist, and US singers from US writers. By observing the spatial connections in this snapshot for other views, we can see how well these groups are separated in the bigger views. In the absence of fraction of their friends and followers they group with other major communities, but as their low degree friends and followers are added, they form their own independent communities.

VIII. Hybrid Analysis

In this section we want to observe the differences between mapping the communities to the other view-snapshots with the two aforementioned strategies in the methodology section, Temporal-then-spatial and Spatial-then-temporal. Fig.17 shows the mapping labels for two different types of communities in different views and over time. Fig.17 shows how a community from the first snapshot and the smallest view maps to the other communities in the larger views and next snapshots. Each cell shows the destination community label and its color indicates whether the result of mapping using each of the strategies is different (red) or it stays the same (gray). Fig.17 illustrates how stable the mapping of Taiwanese core community is regardless of the strategy that we consider for mapping. All cell are gray, indicating that the mapping result is the same using each of the strategies.

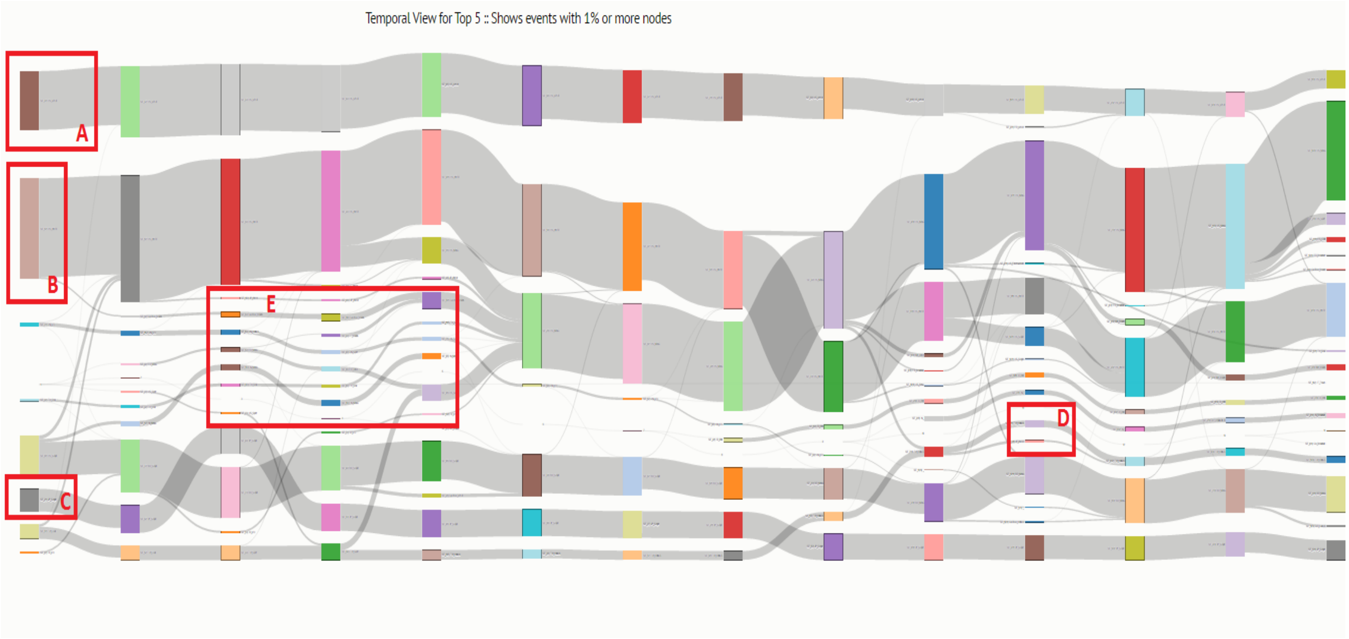
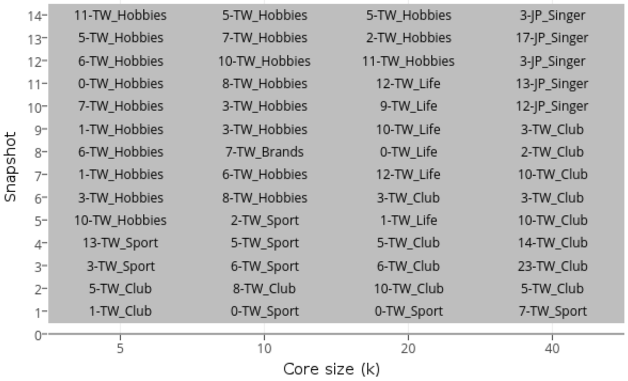


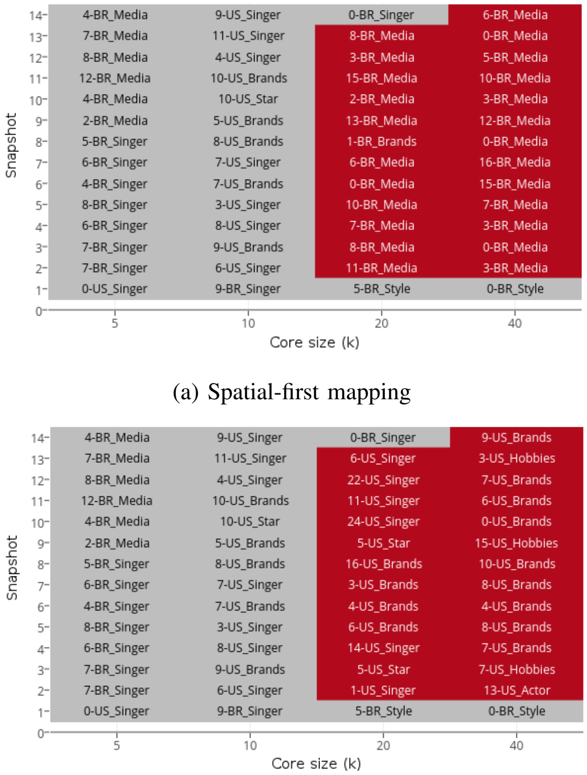
Fig. 16: Temporal analysis of first snapshot, including all node mappings more than one percent of source nodes



Snapshot	5	10	20	40
14	11-TW_Hobbies	5-TW_Hobbies	5-TW_Hobbies	3-JP_Singer
13	5-TW_Hobbies	7-TW_Hobbies	2-TW_Hobbies	17-JP_Singer
12	6-TW_Hobbies	10-TW_Hobbies	11-TW_Hobbies	3-JP_Singer
11	0-TW_Hobbies	8-TW_Hobbies	12-TW_Life	13-JP_Singer
10	7-TW_Hobbies	3-TW_Hobbies	9-TW_Life	12-JP_Singer
9	1-TW_Hobbies	3-TW_Hobbies	10-TW_Life	3-TW_Club
8	6-TW_Hobbies	7-TW_Brands	0-TW_Life	2-TW_Club
7	1-TW_Hobbies	6-TW_Hobbies	12-TW_Life	10-TW_Club
6	3-TW_Hobbies	8-TW_Hobbies	3-TW_Club	3-TW_Club
5	10-TW_Hobbies	2-TW_Sport	1-TW_Life	10-TW_Club
4	13-TW_Sport	5-TW_Sport	5-TW_Club	14-TW_Club
3	3-TW_Sport	6-TW_Sport	6-TW_Club	23-TW_Club
2	5-TW_Club	8-TW_Club	10-TW_Club	5-TW_Club
1	1-TW_Club	0-TW_Sport	0-TW_Sport	7-TW_Sport

Fig. 17: Core community of Taiwanese mapped to the same communities using Temporal-first and Spatial-first strategies.

Fig.18-a and Fig.18-b show that for the second type of communities, such as US_Singer, each strategy might result in different outputs. Starting from the second snapshot and in view top 20, this community maps to different communities in each of the strategies. In general, the more a community merges to others, the more it might map to a different community using these two strategies. In Fig.18-a, US_Singer community maps to BR_Singer in view 10 and then to BR_Style in view 20. Now in temporal view, BR_Style maps to BR_Media. However, in the Fig.18-b, US_Singer first maps to BR_Singer temporally and after two view expansions, it maps to US_Singer in the second snapshot of view 20.



Snapshot	5	10	20	40
14	4-BR_Media	9-US_Singer	0-BR_Singer	6-BR_Media
13	7-BR_Media	11-US_Singer	8-BR_Media	0-BR_Media
12	8-BR_Media	4-US_Singer	3-BR_Media	5-BR_Media
11	12-BR_Media	10-US_Brands	15-BR_Media	10-BR_Media
10	4-BR_Media	10-US_Star	2-BR_Media	3-BR_Media
9	2-BR_Media	5-US_Brands	13-BR_Media	12-BR_Media
8	5-BR_Singer	8-US_Brands	1-BR_Brands	0-BR_Media
7	6-BR_Singer	7-US_Singer	6-BR_Media	16-BR_Media
6	4-BR_Singer	7-US_Brands	0-BR_Media	15-BR_Media
5	8-BR_Singer	3-US_Singer	10-BR_Media	7-BR_Media
4	6-BR_Singer	8-US_Singer	7-BR_Media	3-BR_Media
3	7-BR_Singer	9-US_Brands	8-BR_Media	0-BR_Media
2	7-BR_Singer	6-US_Singer	11-BR_Media	3-BR_Media
1	0-US_Singer	9-BR_Singer	5-BR_Style	0-BR_Style

(a) Spatial-first mapping

Snapshot	5	10	20	40
14	4-BR_Media	9-US_Singer	0-BR_Singer	9-US_Brands
13	7-BR_Media	11-US_Singer	6-US_Singer	3-US_Hobbies
12	8-BR_Media	4-US_Singer	22-US_Singer	7-US_Brands
11	12-BR_Media	10-US_Brands	11-US_Singer	6-US_Brands
10	4-BR_Media	10-US_Star	24-US_Singer	0-US_Brands
9	2-BR_Media	5-US_Brands	5-US_Star	15-US_Hobbies
8	5-BR_Singer	8-US_Brands	16-US_Brands	10-US_Brands
7	6-BR_Singer	7-US_Singer	3-US_Brands	8-US_Brands
6	4-BR_Singer	7-US_Brands	4-US_Brands	4-US_Brands
5	8-BR_Singer	3-US_Singer	6-US_Brands	8-US_Brands
4	6-BR_Singer	8-US_Singer	14-US_Singer	7-US_Brands
3	7-BR_Singer	9-US_Brands	5-US_Star	7-US_Hobbies
2	7-BR_Singer	6-US_Singer	1-US_Singer	13-US_Actor
1	0-US_Singer	9-BR_Singer	5-BR_Style	0-BR_Style

(b) Temporal-first mapping

Fig. 18: US_Singer community mapped to different communities using Temporal-first and Spatial-first strategies.

Therefore, as we analyze the evolution of communities in the graph, we should consider that the starting point in both temporal and spatial angles matters. In the next step, the strategy that we pick to map communities can change the output results. However, there are other type of communities that regardless of starting point or the mapping strategy, map to the same set of communities.

IX. Conclusions

In this study we characterize structural properties of Google+'s network. Having 14 snapshots of this large OSN, we propose a method to capture the evolution patterns of the network through multiple views of high in-degree nodes (core nodes). We show how core nodes form tightly connected communities, which shows strong social cohesion. Then, we characterize the evolution patterns of core communities as we expand the view (Spatial analysis) and over time (Temporal analysis). Different evolution patterns, intra and inter-connectivity of individual core communities, and some of the causes for the observations are discussed in this study. This approach enables us the report the micro level evolution trends of a large OSN. We plan to compare different social networks using the same methodology and also conduct more in-depth analysis on the core communities as manageable units of analysis.

References

[1] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(4):16, 2009.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 44–54, 2006.

[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):34–46, 2008.

[4] P. Bródka, S. Saganowski, and P. Kazienko. Ged: the method for group evolution discovery in social networks. *Social Network Analysis and Mining*, 3(1):1–14, 2013.

[5] N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 131–144.

[6] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas. Google+ or google-?: Dissecting the evolution of the new osn in its first year. In *Proceedings of the 22Nd International Conference on World Wide Web*, pages 483–494, New York, USA, 2013.

[7] D. Greene, D. Doyle, and P. Cunningham. Tracking the evolution of communities in dynamic social networks. In *Advances in social networks analysis and mining (ASONAM), international conference on*, pages 176–183, 2010.

[8] P. Lee, L. V. Lakshmanan, and E. E. Milios. Incremental cluster evolution tracking from highly dynamic network data. In *Data Engineering (ICDE), IEEE 30th International Conference on*, pages 3–14, 2014.

[9] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008.

[10] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[12] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[13] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti. General optimization technique for high-quality community detection in complex networks. *Physical Review E*, 90(1):012811, 2014.

[14] SocialBakers. Socialbakers is the most popular provider of social media analytic tools, statistics and metrics for facebook, twitter, google plus and youtube. [Online; accessed 19-May-2016].

[15] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni. Research on online social networks: Time to face the real challenges. *SIGMETRICS Perform. Eval. Rev.*, 37(3):49–54, 2010.

[16] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao. Multi-scale dynamics in a massive online social network. In *Proceedings of the ACM conference on Internet measurement conference*, pages 171–184, 2012.

[17] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1007–1016, 2009.