DYNAMIC USER-LEVEL AFFECT ANALYSIS IN SOCIAL NETWORKS

[#1]N.ANJALI, M.Tech Student,

[#2]Dr.S.PRABAHARAN, Associate Professor,

[#3]Dr.M.SUJATHA, Associate Professor,

Department Of CSE,

JYOTHISHMATHI INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR T.S.INDIA.

ABSTRACT: Data confidentiality policies at major social network providers have severely limited researchers' access to large scale datasets. The biggest impact has been on the study of network dynamics, where researchers have studied citation graphs and content-sharing networks, but few have analyzed detailed dynamics in the massive social networks that dominate the web today. In this paper, we present results of analyzing detailed dynamics in a large Chinese social network, covering a period of 2 years when the network grew from its first user to 19 million users and 199 million edges. Rather than validate a single model of network dynamics, we analyze dynamics at different granularities (per-user, per community, and network-wide) to determine how much, if any, users are influenced by dynamics processes at different scales. We observe independent predictable processes at each level, and find that the growth of communities has moderate and sustained impact on users. In contrast, we find that significant events such as network merge events have a strong but short-lived impact on users, and they are quickly eclipsed by the continuous arrival of new users.

Keywords: *Dynamic Graphs, Online Social Networks.*

I. INTRODUCTION

A number of interrelated processes drive dynamics in social networks. A deeper understanding of these processes can allow us to better model and predict structure and dynamics in social networks. In turn, improved models and predictors have numerous practical implications on the design of infrastructure, applications, and security mechanisms for social networks. Details of these dynamic processes are best studied in the context of today's massive Online Social Networks (OSNs), e.g. Facebook [38], LinkedIn [24], and Renren [13]. Unfortunately, the providers of large social networks generally consider their dynamic network data to be trade secrets, and have few incentives to make such data available for research. Instead, studies have analyzed citation networks [22], content sharing networks [18], and high level statistics of social networks [1]. Others [21, 26, 10] sought to validate generative models such as preferential attachment (PA) [5].

Our goal is to better understand in detail the evolutionary dynamics in a social network.

This includes not only the initial growth process during a social network's formation, but also the ongoing dynamics afterwards, as the network matures. Much of the prior work in this area, including generative graph models and efforts to validate them [5, 21, 26, 10], has focused on capturing network dynamics as a single process. In contrast, we are interested in the question "how are individual user dynamics influenced by processes at different scales?" How much are the dynamics of users influenced by external forces and events, such as the activities of friends in communities they belong to, or by large-scale events that occur at the network level? In this work, we explore these questions empirically through a detailed analysis of social network dynamics at multiple scales: at the individual user level, at the level of user communities, and at the global network level. We study a dynamic graph, i.e. a sequence of detailed time stamped events that capture the ongoing growth of a large Chinese online social network. With over 220 million users, it is the largest social network in China, and provides functionality similar to Face book. We focus our analysis on first two years of its growth, from its first user in November 2005, to December 2007 when it had over 19 million members. This captures the network's initial burst of growth, as well as a period of more sustained growth and evolution. Our anonym zed data includes timestamps of all events, including the creation of 19 million user accounts and 199 million edges.

This dataset is notable because of three features: its scale, the absolute time associated with each event, and a rare network merge event, when the network merged with its largest competitor in December 2006, effectively doubling its size from 600K users to 1.3 million users in a single day.

Our analysis of network dynamics in this dataset focuses on three different levels of granularity: nodes, communities, and networks. At each level, we search for evidence of impact on user behavior. Along the way, we also make a number of intriguing observations about dynamic processes in network communities and network-wide events.

Individual Nodes. The creation of links between individual users has been studied in a number of contexts, and is long believed to be driven by generative models based on the principle of preferential attachment, i.e. users prefer to connect to nodes with higher degree [5]. Our goal is to extend the analysis of this model with respect to two new dimensions. First, preferential attachment defines how a sequence of edges are created in logical order, but how do node dynamics correlate with absolute time? Second, does the strength of the preferential attachment model strengthen or weaken as the network grows in scale and matures?

Communities. Intuitively, the behavior of a user is likely to be significantly impacted by the actions of her friends in the network. This has been previously observed in offline social networks [39]. Our goal is to empirically determine if user activity at the level of communities has a real impact on individual users. To do so, we first implement a way to define and track the evolution of user communities over time. We use the Louvain algorithm [6] to detect communities, track the emergence and dissolution of communities over time, and quantify the correlation of user behavior to the lifetime, size, and activity level of the communities they belong to.

Networks. Finally, we wish to quantify the impact, if any, of network-level events on individual user behavior. By network-level events, we refer to unusual events that affect the entire network, such as the merging of two distinct social networks recorded in our dataset. We analyze user data before and after the merger of our social network and its competitor, and quantify the impact of different factors on user behavior, including duplicate accounts, and user's edge creation preferences over time.

*Key Findings***.** Our analysis produces several significant findings. First, we find that nodes (users) are most active in building links (friendships) shortly after joining the network. As the network matures, however, we find that new edge creation is increasingly dominated by existing nodes in the system, even though new node arrivals is keeping pace with network growth. Second, we find that influence of the preferential attachment model weakens over time, perhaps reflecting the reduced visibility of each node over time. As the network grows in size, users are less likely to be aware of high degree nodes in the network, and more likely to obey the preferential model with users within a limited neighbourhood. Third, at the level of user communities, we find using the Louvain algorithm that users in large communities are more active in creating friends and stay active for a longer time. In addition, we found that a combination of community structural features can predict the short-term "death" of a community with more than 75% accuracy. Finally, in our analysis of the network merge event, we use user activity to identify duplicate accounts across the networks. Aside from duplicate accounts, we find that the network merge event has a distinct short-term impact on user activity patterns. Users generate a high burst in edge creation, but the cross-network activity fades and quickly become dominated by edge creation generated by new users. Overall, this quickly reduces average distance between the two networks and melds them into a single monolithic network.

II. RELATED WORK

A lot of work in this area concerns the users' perception of privacy in online social networks. According to Acquisti and Gross, users reveal a lot of information on their websites even when they are not very knowledgeable about the security features[1]. Similar concerns are expressed by Edwards and Brown while discussing the threats of default settings in social networking websites[2]. Krasnova et al. looked into the motivations for this kind of user behaviour and are of the view that users disclose information about them for initiating and maintaining relationships and they trust the platform providers. While Dwyer and Hiltz say that users trust the site and many of them also extend the online relationships beyond social networking[3], Krasnova et al. have shown that in wake of social and organizational threats, users withdraw from disclosing information about themselves or can resort to providing false information. Williams et al. are of the view that users' behaviour in the social networks exposes them to various risks. Their study shows that users of different age group behave differently in terms of disclosing personal information online[4]. Stephan Weiss says that threats to privacy differ widely among individuals and are context-specific[5]. All these studies present privacy requirements for the users. However, opinions are divided among how privacy can be ensured. Edwards and Brown believe that technology and law together can provide data protection[2] whereas, James Grimmelmann says that technical control won't work in the case of social networks[6]. Social networking has also gained the attention of administrative and legal organizations.

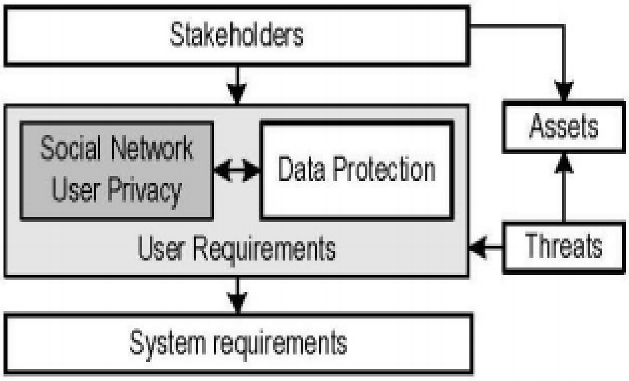


Fig. 1. Privacy requirements in social networks

The Trans-Atlantic Consumer Dialogue, European Network and Information Security Agency (ENISA) and the Data Protection Working Party of the European Commission recently came up with recommendations and opinions in this matter[7-9]. These directives cover various legal and social aspects of privacy and consider the social networking service (SNS) provider responsible for protecting the rights of users in online social networks. These studies are different from our work because we present the privacy requirements from more than one perspectives. While our work builds upon the threats identified by them, we extend the privacy requirements to express them vis-`a-vis system requirements. Also, the thorough classification of all types and aspects of data in social networks is an important contribution. To the best of our knowledge, we did not find any existing work that takes a data-centric approach on this issue. An overview of our work in this paper is shown in Figure 1 where the portions in gray are the focus of existing works.

III. NETWORK LEVEL ANALYSIS

We begin our study by first describing the dataset, and performing some basic analysis to understand the impact of network dynamics on first order graph metrics. Our data is an anonymized stream of time stamped events shared with us by a large Chinese social network, whose functionality is similar to those of Facebook, Google+ and Orkut. Our basic measurements in this section set the context for the analysis of more detailed metrics in later sections.

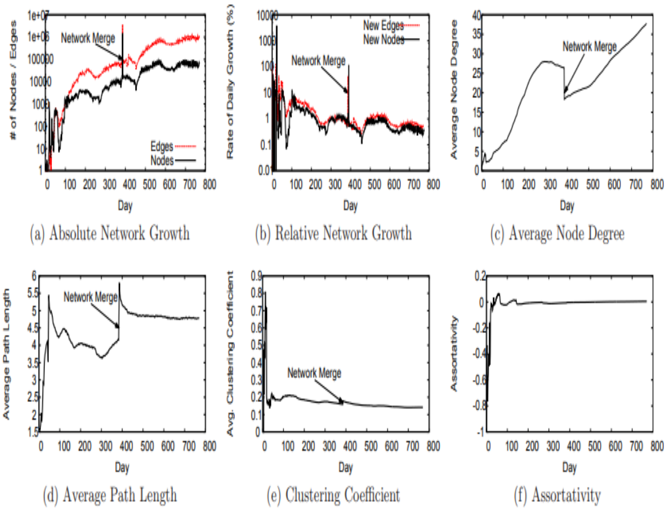
Dataset of Dynamics in a Massive Social Network. The first edge in our large social network was created on November 21, 2005. The social network was originally built as a communication tool for college students, but expanded beyond schools in November 2007.

Our anonymized dataset encompasses the time stamped creation events of all users and edges in the social network. The dataset covers more than 2 years, starting on November 21, 2005 and ending December 31, 2007. In all, the dataset includes the creation times of 19,413,375 nodes and 199,563,976 edges. To perform detailed analysis on the social graph, we produce 771 graphs representing daily static snapshots from the time stamped event stream. Note that in this paper, we will use the term node to mean an OSN user and edge to mean a friendship link.

An unusual event happened on December 12, 2006, when our network merged with a second, competing online social network that was created in April 2006. On the merge date, our social network had 624K users with 8.2 million social links, and the second online social network had 670K users with 3 million social links. Wherever possible, we treat the merge as an external event to minimize its impact on our analysis of network growth. We also present detailed analysis of the network merge event in Section 5. On our network, default user policy limits each user to 1,000 friends. Users may pay a fee in order to increase their friend cap to 2,000. However, prior work by the network has shown that very few users take advantage of such features. We make the same observation about our dataset: the number of users with >1,000 friends is negligibly small.

Network Growth. Figure 1(a) depicts the growth of the large Chinese social network in terms of the number of nodes and edges added each day. Day 0 is November 21, 2005. Overall, the network grows exponentially, which is expected for a social network. However, there are a number of real world events that temporarily slow the growth, and manifest as visible artifacts in Figure 1(a). The two week period starting at day 56 represents the Lunar New Year holiday; a two-month period starting on day 222 accounts for summer vacation; the merge with the competing social network causes a jump in nodes and edges on day 386; additional dips for the lunar new year and summer break are visible starting at days 432 and 587, respectively. In Figure 1(b), we plot daily growth as a normalized ratio of network size from the previous day. It shows that relative growth fluctuates wildly when the network is small, but stabilizes as rapid growth begins to keep rough pace with network size.

Graph Metrics Over Time. We now look at how four key graph metrics change over the lifetime of our data stream, and use them to identify structural changes in the large Chinese social network. We monitor average degree, average path length, average clustering coefficient, and assortativity.



(a) Absolute Network Growth (b) Relative Network Growth (c) Average Node Degree

(d) Average Path Length (e) Clustering Coefficient (f) Assortativity

Figure 2: Network growth over time, and its impact on four important graph metrics.

Average Degree. As shown in Figure 1(c), average node degree grows for much of our observed time period, because the creation of edges between nodes out paces the introduction of new users to the network. When we take a closer look, we see that around days 120, 275, 475 and 650, the average degree grows faster. This means that more edges are created around this time period, which happens to match up nicely with the beginning of new academic semesters over multiple years. On day 305, however, a period of rapid growth in users starts to reduce average degree in the network. This comes from a sudden influx of new users due to several successful publicity campaigns by the social network. Next, on day 386 (December 2006), average degree drops suddenly when 670K loosely connected nodes from a competing social network merges with our social network. Average degree resumes steady growth following the event, again showing edge growth out pacing node growth and increasing network densification [22].

Average Path Length. We follow the standard practice of sampling nodes to make path length computation tractable on our large social graphs. We compute the average path length over a sample of 1000 nodes from the SCC for each snapshot, and limit ourselves to computing the metric once every three days. As seen in Figure 1(d), the results are intuitive: path length drops as densification increases (i.e. node degree increases). There is a significant jump when nodes from the second online social network join the large social network on day 386, but the slow drop resumes as densification continues after the merge.

Average Clustering Coefficient. Clustering coefficient is a measure of local density, computed as the ratio of the existing edges between the immediate neighbours of a node over the maximum number of edges possible between them. We plot average clustering coefficient in Figure 1(e). In early stages of network growth (before day 60), the network was very small and contained a large number of small groups with loose connections between them. Groups often formed local cliques or near-cliques, resulting in high clustering coefficients across the network. Once the network grows in size, average clustering coefficient transitions to a smooth curve and decreases slowly. The network merge produces a small jump, since the competing social network had many small clusters of 3 or 4 nodes that boosted average clustering coefficient.

Assortativity. Finally, we plot assortativity in Figure 1(f). Assortativity is the probability of a node to connect to other nodes with similar degree, computed as the Pearson correlation coefficient of degrees of all node pairs. In the early stages of the network, the graph is sparse and dominated by a small number of supermodels connecting to many leaf nodes. This produces a strong negative assortativity that fluctuates and then evens out as the network stabilizes in structure. Assortativity evens out at around 0, meaning nodes in our network have no discernible inclination to be friends with nodes of similar or different degree.

Summary. We observe that the high-level structure of our network solidifies very quickly. Several key properties stabilize after the first 2 months, with others establishing a consistent trend after 100 days. While the notable network merge with a second, competing social network introduces significant changes to network properties, the effects quickly fade with time and continued influx of new users to the merged network.

IV. COMMUNITY EVOLUTION

In online social networks, communities can be defined as groups of densely connected nodes based on network structure. More specifically, they are groups of nodes where more edges connect nodes in the same community than edges between different communities [29]. Note that these are implicit groups based on structure, and not explicit groups that a user might join or leave. Communities effectively capture "neighborhoods" in the social network. As a result, we believe they represent the best abstraction with which to measure the influence of social neighborhoods on user dynamics. We ask the question, "how do today's social network communities influence their individual members in terms of edge creation dynamics?" To answer our question, we must first develop a method to scalably identify and track communities as they form, evolve, and dissolve in a dynamic network. There is ample prior work on community detection in static graphs [29, 7, 37, 6]. More recent work has developed several algorithms for tracking dynamic communities across consecutive graph snapshots [17, 32, 23, 35, 34]. Some of these techniques are limited in scale by computational cost, others require external information to locate communities across snapshots of the network.

In the remainder of this section, we describe our technique for scalably identifying and tracking communities over time based on network structure. We then present our findings on community dynamics in our social network, including community formation, dissolution, merging, and splitting. Finally, we analyze community-level dynamics, and use our detected communities to quantify the correlation between node and community-level dynamics. To make computation tractable across our large dataset, we choose a modified Louvain algorithm to produce the large majority of our results. To ensure that our choice of community detection algorithm does not significantly bias our results, we validate a portion of our findings using a second community detection algorithm that does not rely on modularity.

4.1 Tracking Communities over Time

Tracking communities in the presence of network dynamics is a critical step in our analysis of network dynamics at different scales. Prior work proved that dynamic community tracking is an NP-hard problem [35]. Current dynamic community tracking algorithms [17, 32, 23, 35, 34, 11] are approximation algorithms that "track" a community over multiple snapshots based on overlap with an incarnation in a previous snapshot. For scalability and efficiency, we use the similarity-based community tracking mechanism [11] for our analysis. In this section, we first introduce background on community detection algorithms and related definitions.

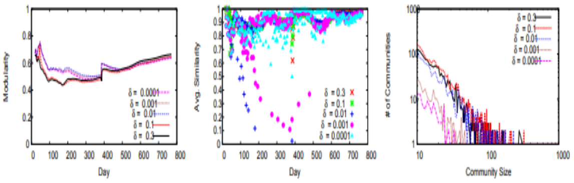
Then, we briefly describe our mechanism, which is a modified version of [11] that provides tighter community tracking across snapshots using the incremental version of the Louvain algorithm [6]. At a high level, we use incremental Louvain to detect and track communities over snapshots, and use community similarity to determine when and how communities have evolved.

Similarity-based Community Tracking.

Louvain [6] is a scalable community detection algorithm that significantly improves both modularity and efficiency using greedy local modularity optimization. It uses a bottom up approach that iteratively groups nodes and communities together, and migrates nodes between communities until the improvement to modularity falls below a threshold δ. To the best of our knowledge, Louvain is the only community detection algorithm that scale to graphs with tens of millions of nodes1 . Our approach leverages the fact that Louvain can be run in incremental mode, where communities from the current snapshot are used to bootstrap the initial assignments in the next snapshot. Given how sensitive community detection is to even small changes in modularity, this approach enables more accurate tracking of communities by providing a strong explicit tie between snapshots. Finally, we follow the lead of [11], and track communities over time by computing the similarity between communities. Similarity is quantified as community overlap and is computed using set intersection via the Jaccard coefficient.

Community Evolution Events.

Using similarity to track communities allows us to detect major community events, including their birth, death, merges, and splits. We define a community A splits at snapshot i when A is the highest correlated community to at least two communities B and C at snapshot i + 1. When at least two communities A and B at snapshot i contribute most of their nodes to community C at snapshot i + 1, A and B have merged. When a community A splits into multiple communities X1, X2...Xn, we designate Xj as the updated A in the new snapshot, where Xj is the new community who shares the highest similarity with A. We say that all other communities in the set were “born” in the new snapshot. Similarly, if multiple communities merge into a single community A, we consider A to have evolved from the community that it shared the highest similarity with. All other communities are considered to have “died” in the snapshot.



(a) Modularity (b) Average Community Similarity (c) Community Size Distribution on Day

Figure 3: Tracking communities over time and the impact of δ. (a) The value of modularity always stays above 0.4, indicating a strong community structure. The choice of δ has minimum impact, and δ = 0.01 is sensitive enough to detect communities. (b) The value of average similarity over time at different δ values. Small δ values like 0.0001 and 0.001 produce less robust results. (c) The distribution of community size observed on Day 602. The algorithm is insensitive to the choice of δ once δ ≥ 0.01. The same conclusion applies to other snapshots.

Sensitivity Analysis. We run the Louvain algorithm on our dynamic graph snapshots generated every 3 days. We start from Day 20, when the network is large enough (64 nodes) to support communities, and only consider communities larger than 10 nodes to avoid small cliques. We scale δ between 0.0001 and 0.3, and plot the resulting modularity and average similarity in Figure 4. As shown in Figure 4(a), in all snapshots the modularity for all thresholds is more than 0.4. According to prior work [20], modularity ≥ 0.3 indicates that our social network has significant community structure. As expected, a threshold around 0.01 is sensitive enough for Louvain to produce communities with good modularity. Note that the big jump in modularity on Day 386 is due to the network merge event. Figure 4(b) shows that thresholds 0.0001 and 0.001 produce lower values of average similarity (i.e. they are less robust and more sensitive) compared to higher thresholds between 0.1 and 0.3. Thus, Louvain with δ > 0.01 generates relatively good stability of communities between snapshots. Lastly, we examine whether detected communities are highly sensitive to the choice of δ. As an example, Figure 4(c) plots the distribution of community sizes observed on Day 602. The conclusion from this figure is that once the threshold exceeds 0.01, the impact of δ on community size is reduced to a minimum. The same conclusion applies to other snapshots as well. Based on the results in Figure 4, we repeat the Louvain algorithm within a finer threshold range of 0.01 to 0.1. We find that a threshold value of 0.04 provides the best balance between high modularity and similarity. We use δ = 0.04 to track and measure dynamic communities in the rest of our analysis on the dataset.

4.4 Impact of Community on Users

To understand how communities impact users’ activity, we compare edge creation behaviors of users inside communities to those outside of any community. Overall, our results show that community users score higher on all dimensions of activity measures, confirming the positive influence of community on users.

Edge Inter-arrival Time. Figure 7(a) plots the CDF of edge inter-arrival times for community and noncommunity users. We observe that users within different communities display similar edge inter-arrival statistics, and merge their results into a single CDF curve for clarity. The considerable distance between the two curves confirms that community users are more enthusiastic in expanding their social connections than non-community users.

User Lifetime. Next, we examine how long users stay active after joining the network, and whether engagement in a community drives up a user's activity span. We define a user i's lifetime as the gap between the time i builds her last edge and the time i joins the network. Figure 7(b) plots the CDF of user lifetime for users in different size communities as well as non-community users. [x, y] represents communities of size between x and y. We find that the lifetime distribution depends heavily on the size of the community. The larger the community is, the longer its constituent user's lifetimes are. Compared to noncommunity users, users engaging in a community tend to stay active for a longer period of time. This confirms the positive impact of community on users.

In-Degree Ratio. Although we know that communities have more edges inside than outside statistically, we want to quantify how each user within each community connect to each other. We compute each user's in-degree ratio, i.e. the ratio of her edge count within her community to her degree. Figure 7(c) shows the CDF of the in-degree ratio for users in communities of different sizes. We observe that users in larger communities have a larger in-degree ratio, indicating that they form a greater percentage of edges within their own community. In particular, 11-30% of nodes only interact with peers in their own communities. These results show that like offline communities, online social communities also encourage users to interact "locally" with peers sharing mutual interests.

V. CONCLUSION

This work presents a detailed analysis of user dynamics in a large Chinese online social network, using a dataset that covers the creation of 19 million users and 199 million edges over a 25-month period. More specifically, we focus on analyzing edge dynamics at different levels of scale, including dynamics at the level of individual users, dynamics involving the merge and split of communities, and dynamics involving the merging of two independent online social networks. Our analysis produced a number of interesting findings of dynamics at different scales. First, at the individual node level, we found that the preferential attachment model gradually weakens in impact as the network grows and matures. In fact, edge creation in general becomes increasingly driven by connections between existing nodes as the network matures, even as node growth keeps pace with the growth in overall network size. Second, at the community level, we use an incremental version of the popular Louvain community detection algorithm to track communities across snapshots. We empirically analyze the birth, growth, and death of communities across merge and split events, and show that community merges can be predicted with reasonable accuracy using structural features and dynamic metrics such as acceleration in community size. Finally, we analyze detailed dynamics following a unique event merging two comparably sized social networks, and observe that its impact, while significant in the short term, quickly fades with the constant arrival of new nodes to the system.

While our results from this network may not generalize to all social networks, our analysis provides a template for understanding the dynamic processes that are active at different scales in many complex networks. A significant takeaway from our work is that the actions of individual users are not only driven by dynamic processes at the node-level, but are also significantly influenced by events at the community and network levels. A comprehensive understanding or model of an evolving network must account for changes at the network and community levels and their impact on individual users.

REFERENCES

[1] Ahn, Y., Han, S., Kwak, H., Moon, S., and Jeong, H. Analysis of topological characteristics of huge online social networking services. In Proc of WWW (2007).

[2] Akoglu, L., McGlohon, M., and Faloutsos, C. RTM: Laws and a recursive generator for weighted time-evolving graphs. In Proc. of ICDM (2008).

[3] Asur, S., Parthasarathy, S., and Ucar, D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM TKDD 3, 4 (2009), 16.

[4] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. Group formation in large social networks: membership, growth, and evolution. In Proc. of KDD (2006).

[5] Barabasi, A., and Albert, R. ´ Emergence of scaling in random networks. Science 286, 5439 (1999), 509.

[6] Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech (2008).

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

[8] Der´enyi, I., Palla, G., and Vicsek, T. Clique percolation in random networks. Physical review letters 94 (2005).

[9] Dunbar, R. Neocortex size as a constraint on group size in primates. Journal of Human Evolution 22, 6 (1992).

[10] Garg, S., Gupta, T., Carlsson, N., and Mahanti, A. Evolution of an online social aggregation network: an empirical study. In Proc. of IMC (2009).

[11] Greene, D., Doyle, D., and Cunningham, P. Tracking the evolution of communities in dynamic social networks. In Proc. of ASONAM (2010).

[12] Guo, L., Tan, E., Chen, S., Zhang, X., and Zhao, Y. Analyzing patterns of user content generation in online social networks. In Proc. of KDD (2009).

[13] Jiang, J., Wilson, C., Wang, X., Huang, P., Sha, W., Dai, Y., and Zhao, B. Y. Understanding latent interactions in online social networks. In Proc. of IMC (2010).

[14] Kairam, S., Wang, D., and Leskovec, J. The life and death of online groups: predicting group growth and longevity. In Proc. of WSDM (2012).

[15] Kang, U., McGlohon, M., Akoglu, L., and Faloutsos, C. Patterns on the connected components of terabyte-scale graphs. In Proc. of ICDM (2010).