Multi-scale Dynamics in a Massive Online Social Network

Xiaohan Zhao, Alessandra Sala,
Christo Wilson, Xiao Wang, Sabrina Gaito,
Haitao Zheng, Ben Y. Zhao

Department of Computer Science, UC Santa Barbara
Bell Labs, Ireland, Peking University,
Università degli Studi di Milano

IMC 2012

Introduction

- Effect of dynamic processes at different scales on users

- Data Set
  - Chinese social network
  - 19 M users , 199 M edges
  - 2 years , 771 days , with time stamp of edge and node creation
  - 650 K each , 8.2 edge and 3 M edges, merged.
  - Limited to 1000 friends

- Community level
  - Moderate
  - Sustained
- Network level
  - Strong
  - Short-lived

Scale of Dynamics

- Individual users
  - Link creation between users
  - PA , users prefer to connect to high degree nodes
  - PA as network grows in scale and matures
- Communities
  - Impact of communities on users activities
- Networks
  - Impact of networks on users activities

Main Findings

- shortly after joining => active in building links
- with network growth => link creation dominated by existing nodes

- influence of the preferential attachment model weakens over time
  - reduced visibility of each node over time
  - Size ▲ awareness about high degree nodes ▼
  - PA on limited neighborhood

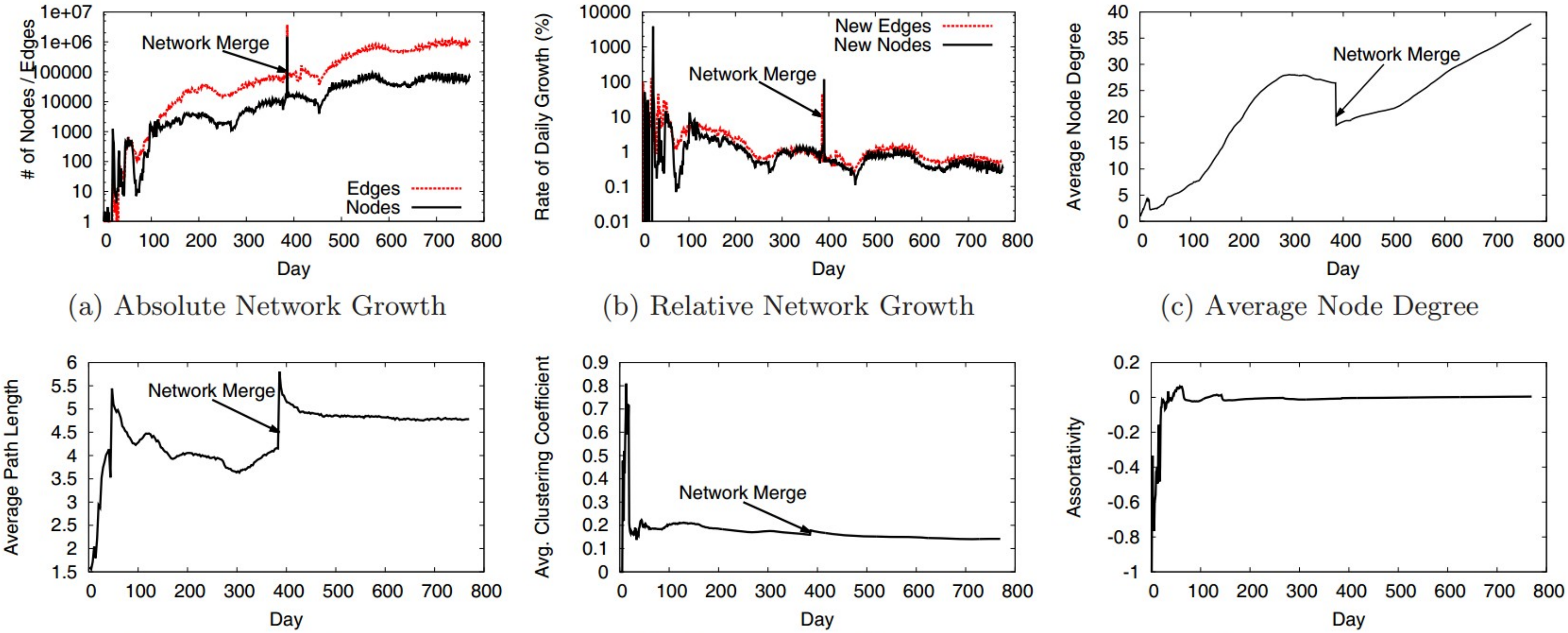
Main Findings

- Users in large Communities
  - More active in link creation
  - Stay active for longer time

- Using some community structure features
  - The death of the community is predictable

- Network event
  - Can increase the edge creation for a short time

Graph metrics over time



(a) Absolute Network Growth

(b) Relative Network Growth

(c) Average Node Degree

(d) Average Path Length

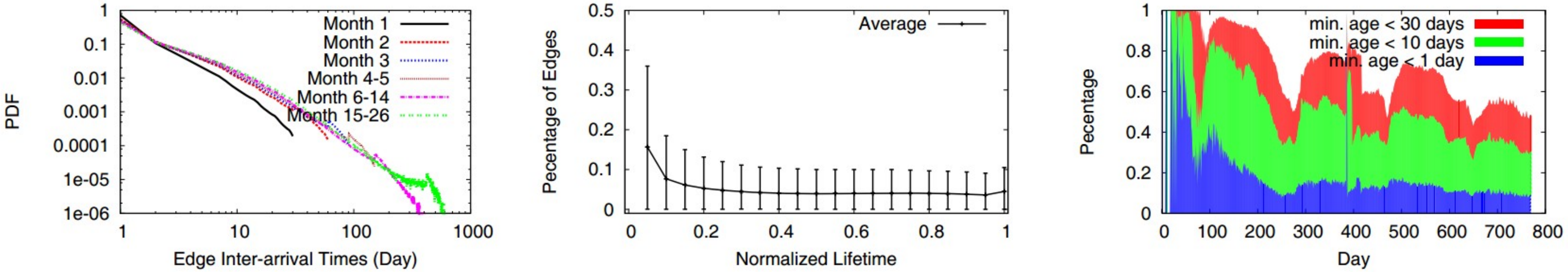
(e) Clustering Coefficient

(f) Assortativity

Figure 1: Network growth over time, and its impact on four important graph metrics.

high-level structure of
our network solidifies very quickly

Dynamics of Edge Creation



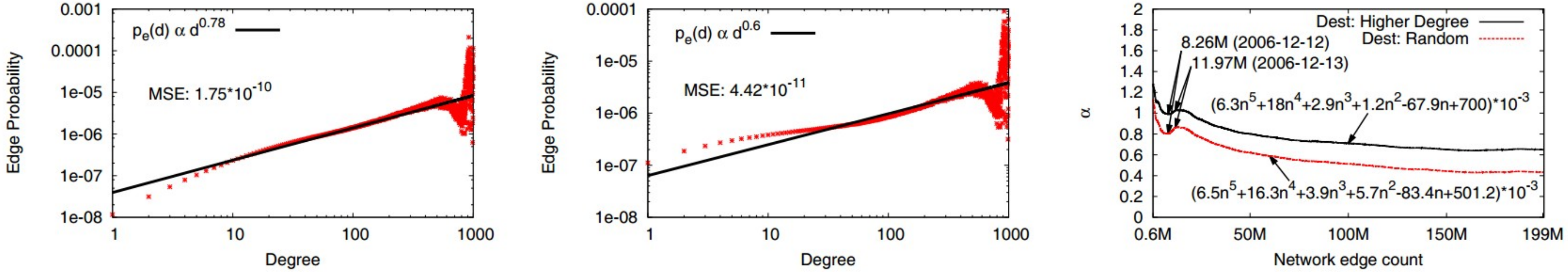
(a) Distribution of Edge Inter-arrival Times

(b) Edge Creation vs. Normalized Lifetime

(c) Node Age and Edge Creation

Figure 2: Time dynamics of edge creation. (a) The probability distribution of the edge inter-arrival times follows a power-law distribution. (b) The normalized activity level over each user's lifetime. Users create most of her friendships early on. (c) The portion of edges created by new nodes each day. When the network is young, new edges are mostly triggered by newly joined nodes. However, as the network matures, the majority of new edges connect older users.

Strength of Preferential Attachment



(a) $p_e(d)$, higher-degree node as destination

(b) $p_e(d)$, random destination selection

(c) Evolution of $\alpha(t)$

Figure 3: (a)-(b) Fitting the measured edge probability $p_e(d)$ with d^α, when our large Chinese social network reaches 57M edges. In (a), $p_e(d)$ is calculated by selecting the higher-degree node as each edge's destination. In (b) the destination is selected randomly. The mean square error (MSE) is very low, confirming the goodness of the fit. (c) As the network grows, α drops from 1.25 to 0.65. It can be approximated by a polynomial function of the network edge count n.

$$p_e(d) = \frac{\Sigma_t\{e_t(u,v) \wedge d_{t-1}(v) = d\}}{\Sigma_t|v : d_{t-1}(v) = d|}$$

Observations

Decreasing of a and two scenario

- Super node based
- Off-line friends based

In a node’s lifetime, edge creation rate is highest shortly after joining the network and decreases over Time.

Observations

- Edge creation in early stages of network growth is
175driven by new node arrivals, but this trend decreases
significantly as the network matures.

- While edge creation follows preferential attachment,
the strength degrades gradually as the network expands
and matures. =>>> we should combine a preferential attachment
component with a randomized attachment component

Community Evolution

- Dynamic community tracking is an NP-hard problem

- Detection:
  - using the incremental version of the Louvain algorithm
    - communities from the current snapshot are used to bootstrap the initial assignments in the next snapshot.

- Tracking:
  - Similarity-based Community

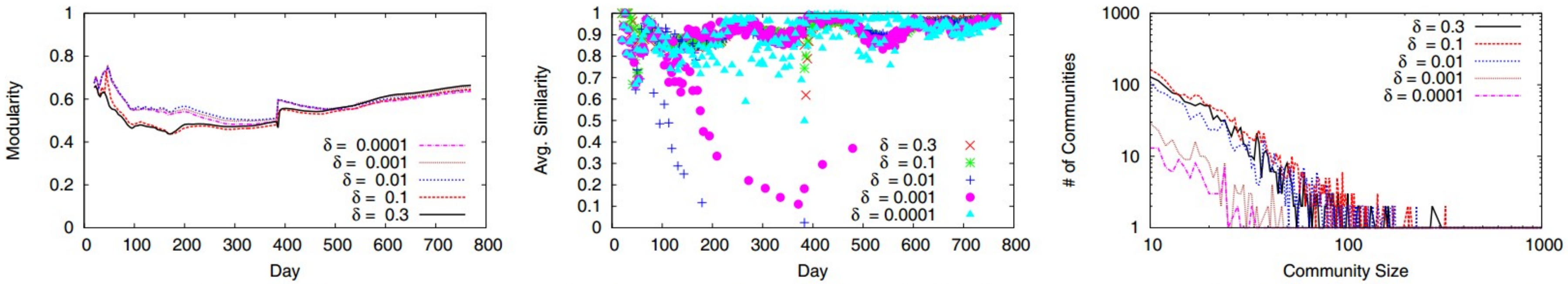
Community Evolution Events

- Birth
  - 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.
- Death
  - 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.

Community Evolution Events

- Merge
  - When at least two communities A and B at snapshot i contribute most of their nodes to community C at snapshot i + 1
- Split
  - A is the highest correlated community to at least two communities B and C at snapshot i + 1.

Sensitivity Analysis



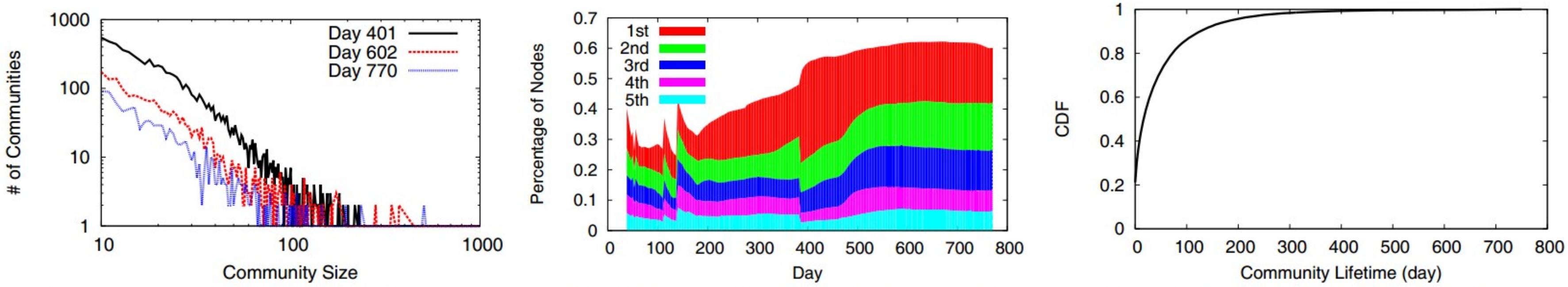
(a) Modularity

(b) Average Community Similarity

(c) Community Size Distribution on Day 602

Figure 4: 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 $\delta = 0.01$ 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 $\delta \geq 0.01$. The same conclusion applies to other snapshots.

Community Statistics Over Time



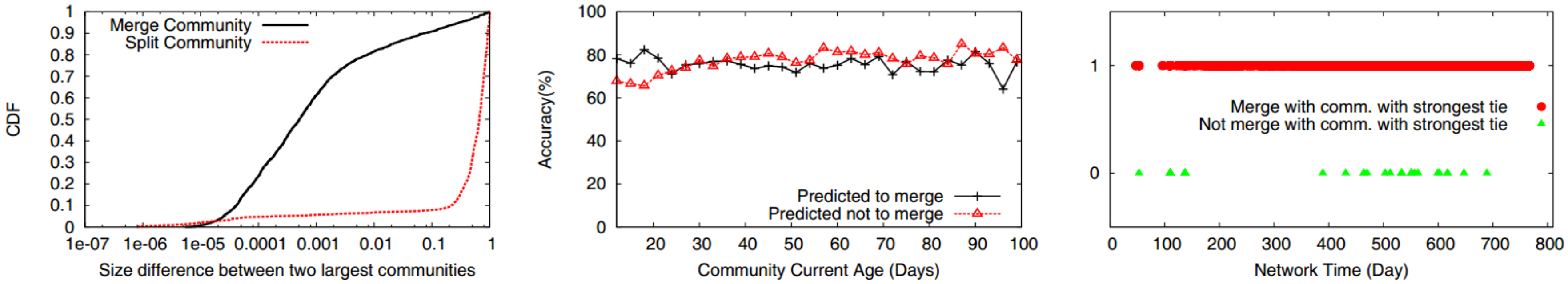
(a) Community Size Distribution

(b) % of Nodes Covered by Top 5 Communities

(c) CDF of Community Lifetime

Figure 5: Analysis on the evolution of communities. (a) Community size distribution on Days 401, 602, and 770. All three lines follow a power-law distribution, and show a gradual trend towards larger communities. (b) The portion of nodes covered by the top 5 communities grows considerably as the network matures. (c) Distribution of community lifetimes shows most communities only stay in the network for a very short time, and are quickly merged into other communities. This indicates a high level of dynamics between communities.

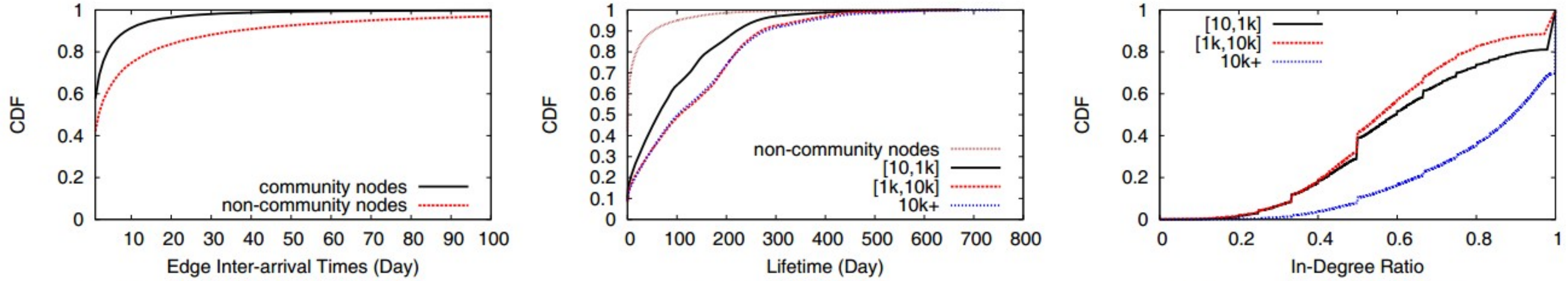
Community Merging and Splitting



(a) Community Size Difference (b) Accuracy on Merge Prediction (c) Patterns of Merged Communities

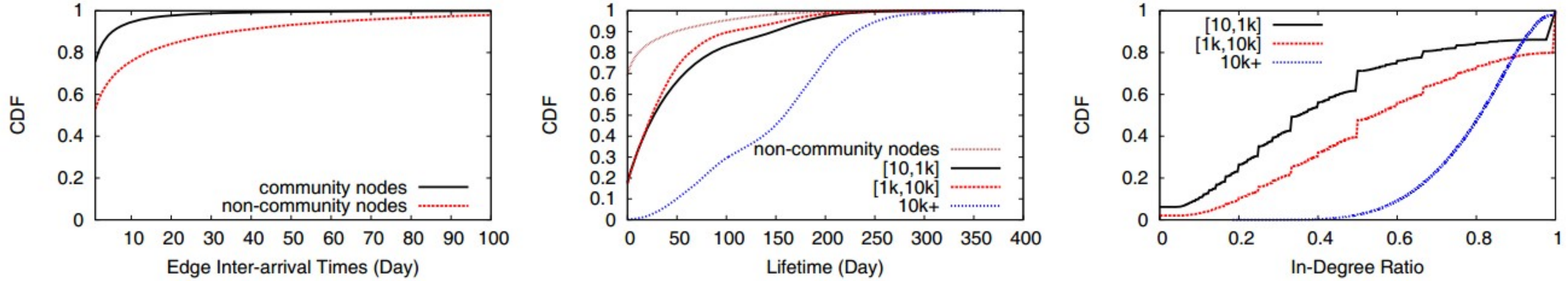
Figure 6: Analysis of community merge and split events. (a) The distribution of the normalized size difference between the largest two components when they split or merge. Small communities always merge into large communities, and a community tends to split into two communities of comparable sizes. (b) The accuracy of our prediction on whether a community will merge with another in the next snapshot. We achieve a reasonably good accuracy of 75%. (c) With very high probability (99%), a community merges with the community that has the most edge connections (or the strongest tie) to itself.

Impact of Community on Users



(a) Edge Inter-arrival Time (b) Node Lifetime (c) Edge In-Degree Ratio

Figure 7: Comparing activity of users inside and outside communities. Community users score higher on all dimensions of activity measures, confirming the positive influence of community on users. (a) Edge inter-arrival time. Community nodes create edges more frequently than non-community nodes. (b) Node lifetime. Community users are grouped by their community sizes. $[x, y]$ represents communities of size between x and y. Community nodes stay active longer than non-community nodes. (c) Community user's in-degree ratio. Nodes in larger communities are more active within their own communities.



(a) Edge Inter-arrival Time (b) Node Lifetime (c) Edge In-Degree Ratio

Figure 8: Verification of results on the impact of communities on users' activities, using the Absolute Potts Model (APM) community detection algorithm. The results from communities detected by APM are consistent with our results using communities detected by our incremental Louvain approach.

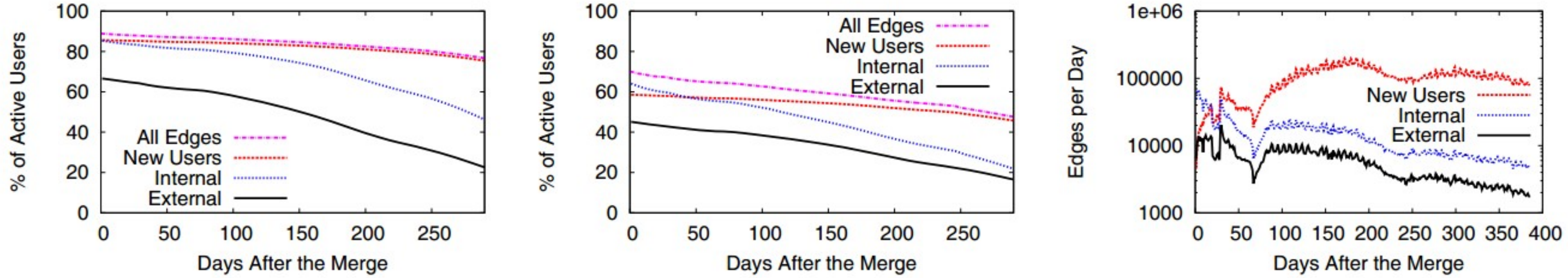
Summary of Results

- Our social network displays a strong community structure, and the size of the communities follows the power-law distribution.

- The majority of communities are short-lived, and within a few days they quickly merge into other larger communities. These merge events can be reliably predicted using structural features and dynamic metrics.

- The membership to a community has significant influence on users' activity. Compared to stand-alone users, community users create edges more frequently, exhibit a longer lifetime, and tend to interact more withpeers in the same community.

MERGING OF TWO OSNS

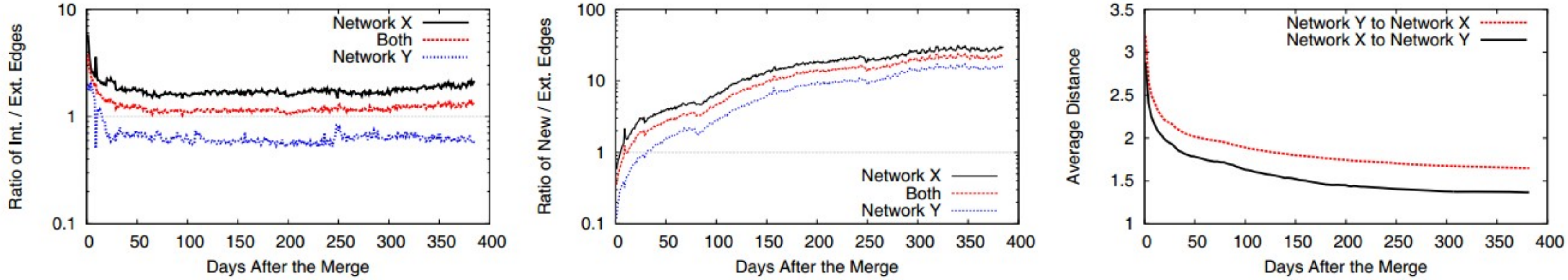


(a) Active users over time for network X

(b) Active users over time for network Y

(c) Edge creation per day

Figure 9: (a)-(b) The number of active users over time. Accounts that are inactive on day 0 after the merge are likely to by discarded, duplicate accounts. Overall user activity declines over time. (c) Number of edges of different types created per day after the merge. Edges to new users quickly become the most popular edge type, although there is a small peak for external edges as well.



(a) Ratio of internal to external edges

(b) Ratio of new to external edges

(c) Distance between 2 OSNs over time

Figure 10: (a) Ratio of internal to external edges over time. Network X users create more edges overall, and are biased towards internal edges, weighting the average upward. (b) Ratio of new to external edges per day. Both networks overwhelmingly prefer edges to new users, although they reach this point at different rates. (c) The average distance in hops between the two OSNs drops over time as more internal and external edges are created. By day 50, the two networks are essentially one large, well connected whole.

Summary of Results

- There were a large number of duplicate accounts between the two networks that become inactive immediately after the merge.

- Edges to new nodes quickly become the driving force behind edge creation.

- Despite user’s preference against external edges, the two networks very quickly merge into a single, well connected graph.