Arrival and Departure Dynamics in Social Networks

Shaomei Wu[*]
Facebook
Menlo Park, CA
shaomei@fb.com

Atish Das Sarma[†]
eBay Research Labs
San Jose, CA
atish.dassarma@gmail.com

Alex Fabrikant
Google Research
Mountain View, CA
fabrikant@google.com

Silvio Lattanzi
Google Research
New York, NY
silviol@google.com

Andrew Tomkins
Google Research
Mountain View, CA
tomkins@google.com

ABSTRACT

In this paper, we consider the natural arrival and departure of users in a social network, and ask whether the dynamics of arrival, which have been studied in some depth, also explain the dynamics of departure, which are not as well studied.

Through study of the DBLP co-authorship network and a large online social network, we show that the dynamics of departure behave differently from the dynamics of formation. In particular, the probability of departure of a user with few friends may be understood most accurately as a function of the raw number of friends who are active. For users with more friends, however, the probability of departure is best predicted by the overall fraction of the user's neighborhood that is active, independent of size. We then study global properties of the subgraphs induced by active and inactive users, and show that active users tend to belong to a core that is densifying and is significantly denser than the inactive users. Further, the inactive set of users exhibit a higher density and lower conductance than the degree distribution alone can explain. These two aspects suggest that nodes at the fringe are more likely to depart and subsequent departures are correlated among neighboring nodes in tightly-knit communities.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: :User/Machine Systems; H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Theory Measurement

Keywords

social networks, user engagement, social graph analysis

1. INTRODUCTION

There has been significant focus on the dynamics of propagation in social networks, ranging from the actual spread of biological infection [19, 9] to such disparate online phenomena as sharing [4], tagging [18], brand adoption [2, 11], recommendations [15], and joining a community [3]. In many cases, such as the decision to become a member or use a product, the story does not end at adoption. Instead, the user may decide at any point to cease using the product, or to depart the community. It is not clear that a decision of this type, to "reverse" a prior socially-mediated decision to adopt, will follow the same dynamics as the original decision to adopt. In this paper, we study this question in the context of arrivals and departures within social networks.

A natural place to look for models of arrivals and departures is the existing literature on the spread of infectious physical disease. These models often include a recovery component[19, 9], which is akin to a reversal of the decision to become infected. Typically, however, this component assumes that an infected user recovers based on properties of the immune system, without reference to any social process. In our case, we are motivated by the metaphor of a user at a party with friends. The user is more likely to attend upon discovering that some number of friends will also attend. If some or all of the friends then opt to depart, whether for a new party or to just go home, then the original user is much more likely to follow suit. Hence, we anticipate to see the network effect in both arrivals and departures.

We begin to address this question with a basic study of the temporal correlation of arrivals and departures, and show that both processes introduce significant correlation; in fact, we show that time intervals between the departure of friends are more tightly distributed than the equivalent distribution of gaps between arrival of friends.

Assuming social influence has some effect on the temporal correlation, we might consider arrival as the propagation of a "join" virus, and departure as simply the propagation of a new virus, in this case representing the decision to cease usage. However, this formulation is at odds in reality. It is

[*] Part of this research was performed while the author was visiting Google Research, Mountain View.

[†] Part of this research was performed while the author was working at Google Research, Mountain View.

WSDM'13, February 4–8, 2013, Rome, Italy.

plausible that seeing one friend join a social network, then two, then three, might impel a user to join, as we see in prior work [3]. However, once a user has two hundred friends, will the departure of one, then two, then three friends have a qualitatively different impact on the user's likelihood to depart? Perhaps like the decision to join, the decision to depart depends more on the number of active friends than the number of inactive friends. Or perhaps departure is a fundamentally different decision that depends on an assessment of the pulse of the neighborhood, captured more accurately by the fraction of friends who remain active. At last, departure can be a process that is driven mostly by exogenous factors rather than social forces.

We study this question in the context of the DBLP coauthorship graph and a large social network, and argue that a hybrid of the social influence models could characterize the observed temporal clustering rather accurately. While the number of active friends is known to be a strong predictor of joining a group, for users with twenty or more friends, overall neighborhood activity, measured by the fraction of friends who remain active, is by far the best predictor of likelihood to depart. Surprisingly, this likelihood is linear in the fraction of active friends throughout almost its entire range, and the linear form is identical in both slope and intercept for several different buckets of neighborhood size. Raw counts of inactive friends have low predictive power, and raw counts of active friends, while stronger, remain weak compared to the overall fraction of active friends. On the other hand, for users with fewer than twenty neighbors, the actual count of active friends remains a strong predictor of likelihood to depart. One interpretation of these findings can be that users with few friends rely heavily on the presence of individual friends, while users with more friends stay for the neighborhood atmosphere - they appear one unit closer to departure by each successive fraction of existing friends observed to depart (this phenomena does not apply for the DBLP dataset where many nodes have very small degree).

From this emerging local picture of behavior, we may then ask how arrival and departure dynamics interact with the global structure of the graph. In particular, we seek to understand where departures happen in the graph. It is possible, for example, that departures tend to occur as high-status users in the core of the graph choose to depart in search of the next big thing. Alternately, it is possible that departure happens first at the "fringes" of the graph, and then spreads inwards from there. We study this problem by computing the average induced degree(or density) and conductance of the subgraphs of active and inactive users through time, and comparing these results to thought experiments in which each node decides independently whether to remain active. These experiments allow us to conclude that a core of active nodes remains at much higher internal density than the set of inactive nodes. We also compare the densities observed against the expected density and conductance under a planted degree constraint model. The results suggest that although the inactive set of nodes densifies, its densification is not just a consequence of the degree distribution, but really a consequence of well-connected cluster of nodes from the fringes departing. We reach the picture that departures happen from the fringes and spread to their immediate neighborhoods, while an internal dense core of active nodes survives. We also build a simple model of network evolution based on affiliation networks that incorporates departures and analyze it theoretically to corroborate some of the observed trends.

2. RELATED WORK

There is a large body of work studying the correlation of activity among friends in online communities (see examples in [1, 3, 7, 18, 22]). Most are forms of diffusion research, built on the premise that user engagement is contagious. As such, a user is more likely to adopt new products or behaviors if his friends do so [3, 15]; and large cascades of behavior can be triggered by the actions of a few individuals [11, 21]. With regard to the effect of local structure on the spread of behavior, empirical work has shown a "diminishing returns" on the correlation of activities among network neighbors[3, 22]. Furthermore, the probability of a user adopting a behavior not only correlates with the number of neighbors who have already adopted, but also with the connectivity among his local neighborhood[3, 22].

In addition to research on local influence, a number of theoretical models have been developed to simulate the growth of social networks over time [17, 13, 16]. By modeling the process of nodes arrival and edges creation, these models can generate graphs with observed evolving macro-level structural properties such as degree distribution, edge densification, and diameters shrinking. Although some work in this domain shows that incorporating random deletion of links can induce degree distribution that better matches the observed power-law pattern[10], most existing empirical work focuses on the growth of networks and the increase of activity. Our paper differs by emphasizing the dynamics of user departure from social networks, and the decline of activity. What leads people to depart from social networks? Is inactivity also contagious? Previous studies on user churn in mobile phone networks suggest the existence of social influence at user's disengagement. Dasgupta *et al.* showed in [8] that the probability of a user churning grows with the number of contacts who already churned. Richter *et al.* studied users in groups based on communication intensity[20], finding a strong correlation between individuals' propensity to churn and the group-level characteristics, especially, the (in)activity of the "leader" of the group. Does the same group effect exist when people disengage from a social network? A big difference here is the visibility of the behavior. In the case of a mobile phone network, leaving the service usually involves notifying existing contacts and signaling them about the disengagement; in online social networks, inactivity is less visible and thus may have less influence on others' behavior. However, the extreme case in which all friends depart suggests that there must be some effect. Given such effect, how would a graph evolve structurally after it stops growing, or starts shrinking? There has been some recent theoretical work on modeling the evolution of network structure in the process of "unraveling"[5]. However, built on top of simple game theory principles, these models are not yet examined by empirical data. To our knowledge, this work is the first to address these questions.

3. DATA

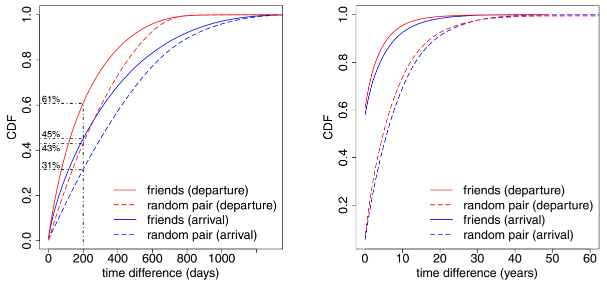
In this paper, we study the dynamics of arrival and departure using a snapshot of the DBLP co-authorship graph and a well-known social network. As previous research [3] showed that the co-author network largely reassembles the dynamics of online social networks in forming individual communities, we are interested at testing whether this similarity persists in the forming and degenerating of the entire social graph. The DBLP snapshot that we consider contains 1 million nodes and around 1.8 millions edges, for each author we store his/her co-authors and the year of the last publication. Furthermore for each author to author edge we also store the year of the first publication. In the rest of the paper we will refer to it as DBLP. The social network dataset we study contains millions of users and over a billion edges. For each user, we have the timestamp of signup and last login, and for each edge, we have the timestamp of edge creation. In the rest of the paper we will refer to this network as SN.

To study the pattern of user arrivals and departures, we first describe each user at each timestamp as either active or inactive, based on his most recent activity time. Given a snapshot of the SN network at time t, we consider a user *inactive* if his last login time is earlier than two months prior to t, and consider a user *active* otherwise[1]. Given a snapshot of the DBLP network at time t, we consider a user *inactive* if he/she has not published any paper in the earlier than five year prior to t, and consider a user *active* otherwise. Note that our results do not depend on the time frame that we used. In fact, they hold for two quite different networks and time frames.

4. ARRIVAL AND DEPARTURE CORRELATION AMONG FRIENDS

In this section, we study the basic properties of arrival and departure. We wish to understand whether users typically arrive and/or depart together in social networks. However, we cannot directly compare gaps between arrivals and departures of friends, as networks are not stationary—consider for example the case of a network that grows very rapidly during a brief period, resulting in a flurry of temporally-proximate arrivals, leading to a mistaken conclusion that arrivals tend to be tightly clustered in time. We must therefore normalize in some way against global rates of arrival and departure, which we do by the following technique. Given a snapshot of the network at time t, we consider two samples of user-pairs, one in which the pair of users are friends, and another in which the pair of users is chosen uniformly from all possible pairs[2]. We then consider the distribution of the gap in arrival time between pairs in the two cases. Differences in these distributions will then highlight temporal correlation of arrivals of friends compared to strangers.

To study departures, we adopt the same technique. We consider only inactive users, and generate again a set of pairs of friends, and another set of pairs chosen uniformly at random. For a fixed time t, we define the last login time of inactive users as their departure time. We pick 1M pairs for each of these four sample groups, and compute the Cumulative Distribution Function (CDF) for these distributions. In Figure 1, we plot the CDF curves, showing the percentage of friends(co-authors)/strangers who joined and left the SN(DBLP) within n days(years) of each other.

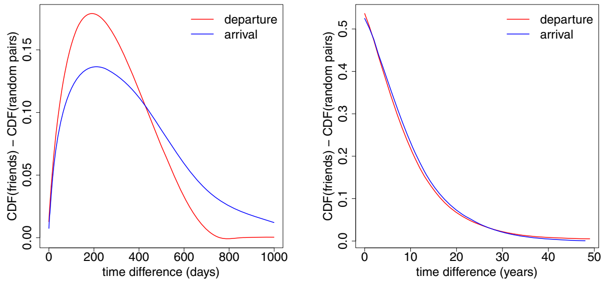


(a) SN (b) DBLP

Figure 1: The CDF curve for the difference in arrival and departure time between friends and random pairs of users.

The CDF for both arrivals and departures of friends lies significantly above the CDF for random pairs, indicating that friends both arrive and depart together, in comparison to the control group of random pairs. As the figure shows, in the case of SN, 43% of random pairs depart within 200 days of one another, while 61% of friends depart within the same period. We find similar pattern in the time interval of arrival - only 31% of random pairs arrive within 200 days, but 45% of friends arrive within the same period. This observation is even more evident in DBLP, where the solid and dashed lines show stronger separation.

To quantify the differences, we plot in Figure 2 the distribution of absolute difference in the CDF values at each time, for arrivals and departures. The correlation of departures in SN is seen to be stronger than the correlation of arrivals, although the two gaps peak around roughly the same value. However, arrivals and departures behave almost identically in DBLP, suggesting the dynamics of arrival and departure are more similar in collaboration networks than in social networks.



(a) SN (b) DBLP

Figure 2: Gap between CDF curves.

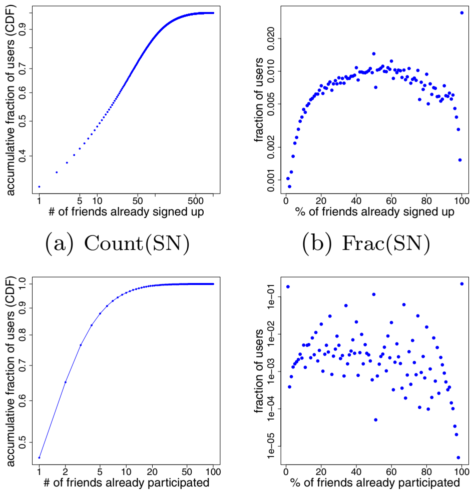
The temporally correlated activities between friends have been observed and modeled before[3, 18, 1], mostly in the form of adopting a new behavior or product within an existing social network. However, given that there is not a single underlying social network that might be accountable for the joining of new users, are those models still applicable in the scenarios of arrival? To answer this question, we will now focus on individual users, and study one's arrival in position of his local neighborhood.

Since we have no information on friendship outside the

[1] As most online social network or social game sites use Monthly-Active-Users(MAU) as a standard way to measure the number of engaged users, we double the cutoff window to 2-month for a more conservative threshold to determine a user has departed.

[2] Technically, it is possible for a random pair to be a pair of friends, however, given the service policy that each user has a rather small upper-bound for the number of friends, the chance of a random pair being friends is negligible.

networks we are studying, we use the eventual set of friends acquired by a user at the snapshot time t to approximate the set of friends he had before arrival, and ask whether those friends join before or after the user. For DBLP, we see the "diminishing returning" curve (Figure 3c) as found in previous research [3]. In SN (Figure 3a), however, the probability of a user signup increases near-linearly as the number of adopted friends increases. The expected fraction of friends joining before the user is 0.5, as the friend network is undirected and each edge contributes one pair in which a joins before b, and one pair with b before a. Thus, for regular graphs (of constant degree), the mean fraction of friends already signed up will be 0.5. The results are shown in Figure 3. True social networks are of course non-regular, and while the distribution of plot (Figure 3b) appears largely symmetrical, there are some outliers. In particular, in SN there are more than 20 times as many users who signed up after all of their eventual friends did, compared to users who signed up before any of their friends. This follows from the many low-degree nodes who join in response to an invitation but do not subsequently engage with the network. In DBLP, the peaks at 0%, 25%, $\frac{1}{3}$, 50%,..., 100% (Figure 3d) can be explained by the substantial number of authors with only one publication, as they have a very small set of co-authors. Although the plot in Figure 3d is largely symmetrical but is slightly skewed to the left. This shows that there are more authors whose collaboration networks grow than ones who stay with their early collaborators. Overall, we posit a weak network effect for new users in both networks, which may not be enough to actively engage users after they sign up.



(a) Count(SN) (b) Frac(SN)

(c) Count(DBLP) (d) Frac(DBLP)

Figure 3: Count and fraction of friends signed-up before user.

Our conclusion from these graphs is that friends tend to arrive and depart together, but at least for SN, departures are more tightly clustered than arrivals. We also find interesting nuances in the dynamics of arrivals, that was not noted previously. In the next section, we will look more closely at departures, learning the how neighborhood activities correspond to one's likelihood to depart.

5. LOCAL NEIGHBORHOODS

Figure 1 shows a strong correlation in arrivals and departures for friends; now we will go beyond single-edges and study how such correlation is presented in the entire neighborhood of a node.

5.1 Dependence on local properties

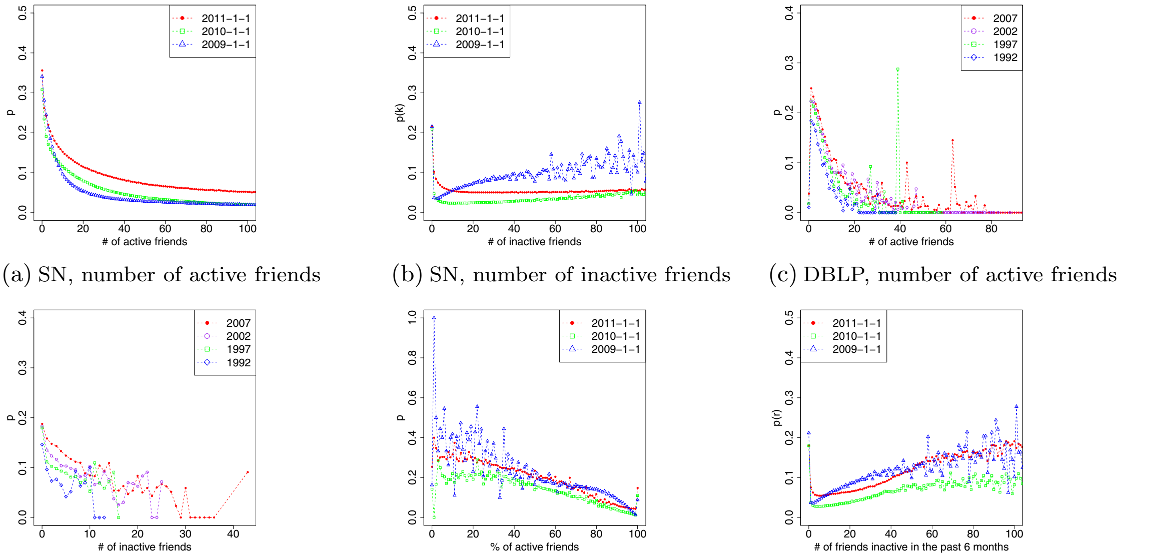
To better understand how a user's departure corresponds to his local community, we look at the probability of a user's departure in relation to the following four properties of the user's neighborhood.

- number of active friends;
- fraction of active friends;
- number of inactive friends;
- number of inactive friends who left in the past 6 months;

We use a similar method as in [3] to calculate the probability of a user becoming inactive, as a function of the number of active friends: we first take two snapshots (t_0, t_1) of the network, three months apart in SN and three years apart in DBLP; we then find all pairs (u, k) such that u is active at the time of first snapshot t_0, and has k friends who are also active at t_0; $p(k)$ is calculated as the fraction of such pairs (u, k) for a given k such that u had left the network at the time of second snapshot t_1. In other words, $p(k)$ is the fraction of active users who left the network in the next three months, given that k friends were active at the first snapshot time. Figure 4a and Figure 4c shows the curves of $p(k)$ at three different t_0. In a similar way, we can fix the fraction of friends f who are active at time t_0, and calculate the probability $p(f)$ of an active user leaving the network as function of f (see Figure 4e). Note that in all figures involving the fraction of active/inactive friends, we exclude all nodes with no friends in SN (around 10% of all active users as of 2011/1/1). Among those users, 35% of them left within three months.

Not surprisingly, Figure 4a, Figure 4c and Figure 4e show that as more and more friends stay active, a user is less and less likely to be inactive. The curve of $p(k)$ (see Figure 4a) also matches very well with what has been seen in other domains [3], exhibiting the "diminishing returns" property. This observation indicates that the marginal gain of having each additional active friend is quite significant for users with a small number of active friends, but rather negligible when a user already has many, say more than 50, active friends. In contrast, in Figure 4e, we do not see such a "diminishing returns" trend, but a steeper, and almost constant rate of decrease in the probability of departure throughout the course when the fraction of active friends increases. This is an interesting observation that has not been previously seen (specifically in various positive influence studies).

To see how the inactivity of the neighborhood determines the departure of a user, we also plot the probability of departure as a function of number of inactive friends, in Figure 4b and Figure 4d. The curves in Figure 4b and Figure 4d show an interesting trend of decreasing slope through time: while the probability of a user departing increases with the growth in the number of inactive friends initially, it becomes more and more insensitive to the value of k in the later curves. This phenomenon is quite intriguing to us: if the departure of friends do have certain predictive power on the departure of the user, as shown in the earlier curves, why is such predictive power diminished so much in the latest years? To answer this question, we note that we are counting the number of inactive friends as prior to the time of each snapshot, but many of them could have been inactive for a long time thus could hardly account for the dynamics of the network



(a) SN, number of active friends (b) SN, number of inactive friends (c) DBLP, number of active friends

(d) DBLP, number of inactive friends (e) SN, percentage of active friends (f) SN, number of recent inactive friends

Figure 4: Probability of departure as function of different local properties. (a) f(active friend count) in SN; (b) f(inactive friend count) in SN; (c) f(active friend count) in DBLP; (d) f(inactive friend count) in DBLP, (e) f(active friend fraction) in SN; (f) f(inactive friends who left in the past 6 months) in SN.

at the snapshot time. Figure 4f confirms this idea, showing that the curves we see in Figure 4b are somewhat misleading - in general, the probability of user's departure constantly grows with the number of friends r who *recently* became inactive (when r is not too small).

5.2 Interaction between local properties

The results of the previous section provide qualitative evidence that an individual's probability of departure is related to the activeness of his neighborhood. However, does that apply to all users? Do the highly connected users act differently than the more marginally connected ones? Is the probability of departure sensitive to the degeneration of neighborhood, or is it a step function that will only drop once there are less than k active friends, as modeled in [5]? To address these issues, we compute the probability of user's departure in SN in relation to the interaction between local properties. Specifically, in Figure 5a, we divide users into three groups based on their degrees, and plot the probability of departure as a function of the number/fraction of active friends, for each group separately. We note that for users with different levels of connectivity in the network, the curves of $p(f)$ (Figure 5a) are qualitatively identical. This result demonstrates again that the fraction of friends who are active has a stronger effect on the probability of an individual's departure, regardless of the size of the user's neighborhood.

In addition, we aggregate users by the fraction of active/inactive friends, and look at how the probability of departure depends on the number of active/inactive friends for each group (see Figure 5). There are two things we note from Figure 5: First, for users with different fractions of inactive friends, there is a big gap between their probabilities of departure - for example, compared to users with less than 10% friends inactive (blue line in Figure 5c), users who have more than 50% friends inactive (red line in Figure 5c) are 10 times more likely to leave as well. Second, once the user is in an inactive part of the neighborhood, the raw count of inactive friends has little effect in determining the probability of the user's departure (green line in Figure 5b). Note that the blue line in Figure 5b is very noisy because there are very few people in a highly obsolete neighborhood but still with a substantial amount of active friends. We still plot it just to be symmetric with Figure 5c.

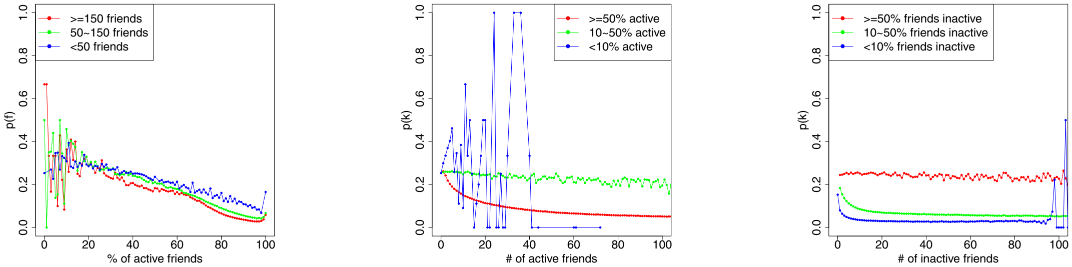
5.3 Predict the departure of user

Given a strong correlation between the probability of a user becoming inactive and the inactivity of his friends, the next question is, can we actually predict individuals's departures based on local properties? In this section, we model the departure of users using simple linear regression models and decision tree classifiers. In particular, we will focus exclusively on SN because we have a richer set of features available.

To start, we formalize our problem as a binary classification task in which class 1 is defined as consisting of those users who were active as of Jan 1st, 2011 (t_0) and departed within two months after t_0, and class 0 is defined as consisting of those who stayed active for two months after t_0. We then randomly sample 500K examples for each class, from all the users who were active at t. Note that in our data, there are 90% negative and only 10% positive cases; our sampling scheme provides a more balanced distribution of examples of both classes.

We extract two sets of local features for each user:

- Neighborhood features. The local structural properties of the user's direct neighborhood, including the number of friends who already departed, the number of friends who are active, the number of friends who departed recently (six months prior to t_0), and the fraction of friends who departed recently.
- Activity features. The properties reflecting user's participation to activities in the network, including the number of contents he received, the number of contents he sent, and the number of status updates.



(a) $p \sim$ percentage of active friends (b) $p \sim$ number of active friends (c) $p \sim$ number of inactive friends

Figure 5: Probability of departure as function of local properties, at different levels of active/inactive friend fraction and friend count (snapshot taken at time t =2011/1/1). in SN

Table 1: Predict user departure with decision tree

Feature	*Accuracy*	*F1 pos*	*ROC area*
Neighborhood	0.694	0.694	0.755
Activity	0.730	0.735	0.801
All	0.755	0.761	0.833

To predict the departure of users, we train a simple decision tree (REPTree) classifier on our examples. Table 1 gives the performance of the classifier with different sets of features under 10-fold cross validation.

Table 1 shows that relying on only local features of individuals, the simple decision tree classifier can predict the departure of user with high accuracy (75% with all features, as compared to 50% for always predicting one class). This result demonstrates a strong connection between user's local properties and the propensity of departure. Moreover, comparing across 3 sets of features, we see that although the activity features are most effective, neighborhood features can also provide rather accurate insights on the departure of users.

6. STRUCTURAL TRENDS IN NETWORK TOPOLOGY

In this section, we explore the overall structural changes that occur in the network as a result of the departure of existing users, as well as the steady arrival of new users. Topological changes have been studied in the context of new nodes arriving but here we pay specific attention to how the global structure changes in the process of the departure or decline of user activities.

To get a sense of the how the structure of the network evolves over time, we first study the distribution of edges among active and inactive nodes. Specially, we look at the edges between active nodes (Figure 6a and Figure 6d), edges between inactive nodes (Figure 6b and Figure 6e), and the edges across active and inactive nodes (Figure 6c and Figure 6f), and plot the ratio between the actual number of edges over the expected value over time. The expected number of edges is computed based on the total number of edges, $|E|$, in the network and the number of nodes in each of the active and inactive sets. The expected number of edges of any type is the expected number of edges when the total $|E|$ edges are placed between randomly chosen pairs of nodes.

To understand the overall structure among the sets of active and inactive nodes, we study the density and conductance of these two sub-networks in the rest of this section. Here the active sub-network consists of the active nodes and the edges among them, and the inactive sub-network is similarly defined.

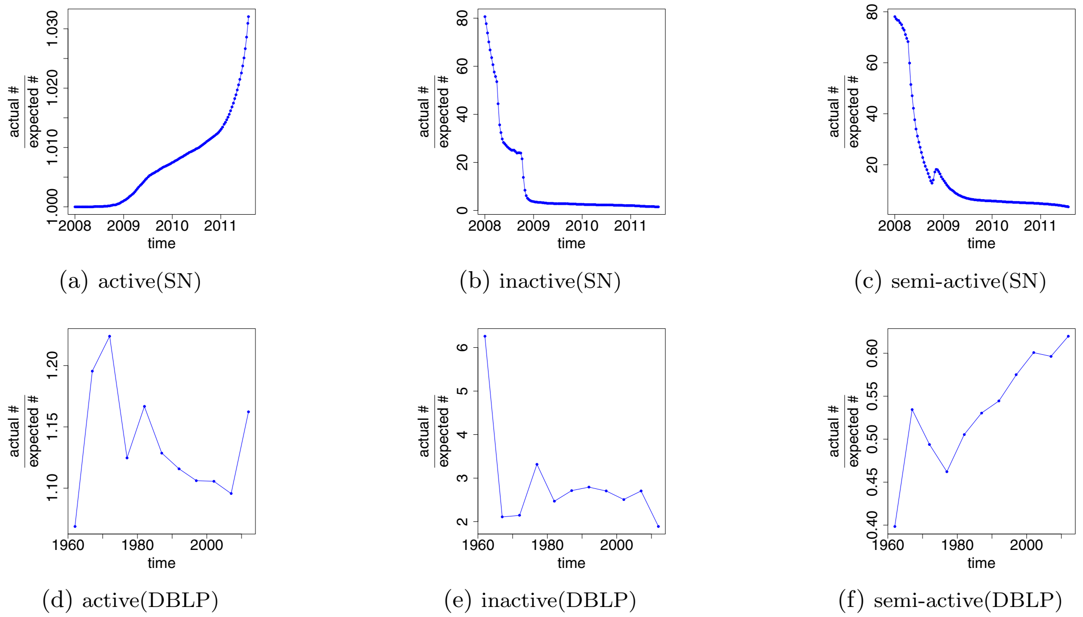
Figure 7 and plots the overall density of the active (7a and 7c) and inactive (7b and 7d) set of nodes, as a function of time. For comparison, we also plot the *expected* densities of the respective sets, as determined by the number of active and inactive nodes and edges and the degree distributions.

We use the definition density of a set of nodes (or average induced degree) used in [6], where the density of a set of nodes is the number of edges between the nodes divided by the number of nodes; i.e. for a set of nodes S, $density(S) = \frac{|E(S,S)|}{|S|}$ (here $E(S,S)$ contains all edges (u,v) such that $u, v \in S$). Therefore, the density of set S is half of the average induced degree of the set of nodes in S. In order to compare the the density we observe for the set of active nodes and the set of inactive nodes, we define an *expected* density for each sub-network. The expected density of the inactive set of nodes could be computed simply as the density of the entire graph times the fraction of inactive nodes.

However, we even use a stronger baseline to see if the trends we observe are a result of a trend more than just that of degrees. Therefore, we compute expected density subject to the overall degree constraints on active and inactive nodes as follows.

Consider each edge as occupying two slots (end points), each slot being in either S_a (the active set of nodes), or S_i (the inactive set of nodes); therefore $S_a \cup S_i = V(G)$. Let the fraction of all these slots that are in S_i be P_i (which is the number of edges going across the active and inactive sub-network plus twice the number of edges in the inactive sub-network); therefore the number of such slots occupied in S_a is $P_a = (1 - P_i)$. Suppose that all the $|E|$ edges were randomly placed in two slots each, with probabilities determined such that in expectation we respect P_i and P_a, then we consider the induced density of this process as the expected density (for respective sub-networks). Notice that this is a more stringent baseline for our comparison. Therefore, an edge is contained in the inactive sub-network with probability P_i^2 and so the expected density of the inactive set is $(|E|P_i^2)/|S_i|$. Similarly the expected density of the active sub-network can be computed.

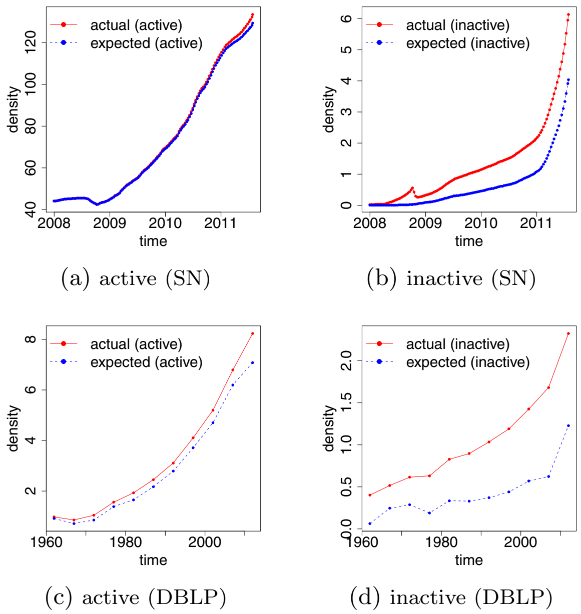
The plots on these densities in Figure 7 shows that the density of the active set $density(S_a)$ increases rapidly with increase in time. However, as shown in the plot on distribution of edges in Figures 6, as the number of edges in the active sub-network continue growing, the density of the ac-



(a) active(SN) (b) inactive(SN) (c) semi-active(SN)

(d) active(DBLP) (e) inactive(DBLP) (f) semi-active(DBLP)

Figure 6: Distribution of edges, indicated by the ratio of actual number of edges over the expected number of edges.



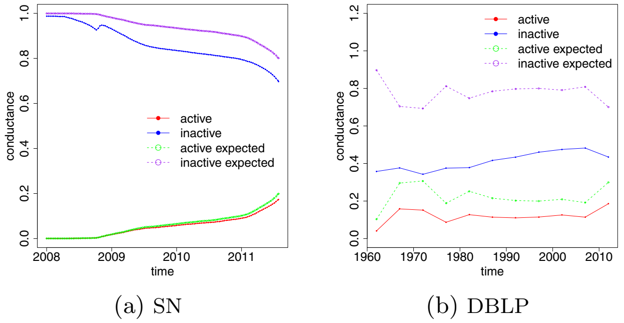
(a) active (SN) (b) inactive (SN)

(c) active (DBLP) (d) inactive (DBLP)

Figure 7: Density of the active and inactive sub-networks

tive sub-network is only marginally higher than its expected density. On contrast, the density of inactive sub-network is significantly higher than the expected density, even conditioned on the degree distribution. This further confirms the fact that departure is correlated across edges, as shown in our local analysis. The nodes that are departing are still probably at the periphery of the network (since the inactive set has much lower density than the active set), but these inactive nodes continue to be internally well-connected because of a higher-than-expected density. This strengthens the evidence from previous sections that a node's likelihood to become inactive is strongly associated with neighboring inactivity.

After studying the connectedness within the active/inactive sub-network separately, we now look at the connection of each sub-network to the rest of social graph, to get a more complete picture about where departures happen and spread in the network. We use conductance to measure the amount of possible connections between different sets of nodes in a network.



(a) SN (b) DBLP

Figure 8: Conductance of the active and inactive sets

Conductance of a set of nodes S, $\phi(s)$, first defined in [12], is measured as $\phi(S) = \frac{|E(S,V(G)\S)|}{|E(S)|}$. Here $E(S, V(G)\S)$ contains all edges (u, v) such that $u \in S, v \notin S$, and $|E(S)| = 2|E(S,S)|+|E(S,V(G)\S)|$. So notice that conductance is always less than 1, and any set with more than half its edges going across to the complement set has a conductance of more than $\frac{1}{3}$. We again measure the conductance of sets S_a and S_i through time and compare with their expected conductances (see Figure 8). The computation of expected conductance is also performed in a similar manner to as described previously for expected density.

We see a similar trend in conductance (Figure 8) as seen for densities. The conductance of the active set of nodes S_a,

$\phi(S_a)$ remains less than the conductance expected for this set. This suggests that there are fewer edges going across from S_a to the inactive set S_i and far more edges within S_a itself, than would be expected. The conductance plots for the set of inactive nodes however is again more contrasting. $\phi(S_i)$ remains far lower than the expected conductance. Nodes that are becoming inactive continue to have many more edges within, than one would expect. This clearly suggests that the inactive set of nodes are influencing neighbors to inactivity. Yet again, the absolute conductance value still suggests that nodes at the periphery of the network are more susceptible to becoming inactive.

The takeaway from these plots are two fold. Firstly, of course, these trends corroborate our findings from the previous sections suggesting that there is a strong correlation among inactivity in neighboring nodes . However, these plots on global measures such as density and conductance also suggest a picture of the evolving network. With the active sub-network's density being much higher than the inactive, and the inactive set showing higher than expected density and lower than average conductance, we are led to believe that nodes in the *core* of the network are much more likely to survive, while nodes at the periphery are more susceptible to departure, probably by a combination of external forces and the neighborhood inactivity.

7. MODELING THE ARRIVAL AND DEPARTURE DYNAMICS

In this section we introduce a simple evolving model that is able to explain formally the densification of the active part of the network. To the best of our knowledge this is the first model that explains the densification of the edges in a network where nodes and edges join and leave the network. Our model is an extension of the Affiliation Networks model introduced in [14]. [3]

First note that the arrival dynamics are already captured by the Affiliation Networks model, where at each step a node is added with constant probability. In order to model the departure probability we introduce in the Affiliation Network model a departure probability of a node, we define the probability of leaving the network for a node v in a step as

$$P[v_{\circ\leadsto}] = f(n) + h(N(v))$$

where $f(n)$ is a function of the size of the network, n, and captures the probability of a node leaving independently of its neighborhood and $h(N(v))$ is the probability of leaving as a function of a node neighborhood, $N(v)$.

In particular we fix $f(n) = \frac{\alpha}{n}$, where α is a constant smaller than 1. Furthermore from figures 4a and 4c we have that $h(N(v)) \propto \left(\frac{1}{|N_{\odot}(v)|}\right)^{\lambda}$, where $N_{\odot}(v)$ is the number of active neighbors of v and α is a constant and $\lambda = g(n)$ where $g(n)$ is function slowly growing with the number of nodes in the network[4]. So we define:

$$P[v_{\circ\leadsto}] = \frac{\alpha}{n} + \left(\frac{1}{|N_{\odot}(v)|}\right)^{\lambda}$$

Here we only consider the dependency on the number of active friends, since in section 5 we showed empirically that this is the more relevant feature. We now present the affiliation network model and our extension to it, and then apply our extension of the model to explain the densification of the active part of the network.

7.1 Our model

We now recall the Affiliation Network model presented in [14] and explain our extension. In the models there are two graphs that evolve at the same time: a bipartite graph between people and their interests and a social network just on the people. The evolution of the two graphs is described in table 2.

In our extension a node v in P in $B(P, I)$ and $G(P, E)$ can also become inactive. More precisely, with probability $P[v_{\circ\leadsto}]$ at each (discrete) timestamp after the first departure time t_d (the time when people start to leave the network), a node becomes inactive with probability $P[v_{\circ\leadsto}]$. Hereafter, we assume that t_d is equal to ϵn, where ϵ is a constant and n is the time at which we analyze the network.

7.2 Densification

In this subsection, we prove that, in our network model, the graph induced by the active nodes densifies even when we allow deactivation of the nodes.

THEOREM 1. *There exist small constant value of λ, α and ϵ for which our model densifies in time with high probability.*

Where with high probability we mean probability bigger than $1 - o(1)$. The detailed proof of this theorem is postponed to the appendix but we sketch the main intuitions here.

PROOF. (Sketch) First we notice that from [14] we have that at time t_d there exist several nodes in I of polynomial degree in the bipartite graph $B(P, I)$ and thus a dense community in $G(P, E)$.

Then we prove that even if part of the network is becoming inactive, there is still a significant number of new user joining the dense community. Furthermore the new nodes will tend to add edges(via preferential attachment) to popular existing community. So we have that even if some nodes are becoming inactive, the dense part of the graph is still growing.

The last observation with the analysis of densification presented in [14] implies that the network is still densifying. □

Note that the main proof's idea are in line with our experimental finding; in fact it suggests that most of the nodes that are leaving the network are in the fringes and not in the core part of the graph (that is composed by the dense communities which continue to grow with time).

Finally it is interesting to notice that the preferential attachment, that does not have a strong clustering structure, do not have the densification property. In fact, the density of the preferential attachment model is also always upper-bounded by a constant when we allow node deletion. Specifically, the density of the preferential attachment model without node deletion is m, where m is the number of edges added by a new node; and it is $\leq m$ for the model with deletions. To prove this, note that we can attribute inactive nodes to inactivation of the m edges that had been added when the node joined the network.

[3] We here study the Affiliation Network model because it gives us a clear way to identify community in the network.

[4] Note that if λ was a constant, all the constant degree nodes would disappear from the networks in a constant number of steps and this is not the case.

$B(P, I)$	$G(P, E)$
Fix integers $c_p, c_i > 0$. Fix $\beta \in (0, 1)$. At time 0, the bipartite graph $B_0(P, I)$ is a simple graph with at least $c_p c_i$ edges.	Fix integers c_p, c_i, s. Fix $\beta \in (0, 1)$. At time 0, $G_0(P, E)$ two vertices have an edge between them for every neighbor in I that they have in common in $B_0(P, I)$.
At time $t > 0$:	At time $t > 0$:
(Evolution of P) With probability β: A new node v is added to P. A node $v_i \in P$ is chosen as *prototypes* for the new node, with probability proportional to its active degree. c_p edges are "copied" from v_i	**(Evolution of P)** With probability β: A new node v is added to P. An edge between v and another node in P is added with probability p for every neighbor that they have in common in $B(P, I)$. The new node also add s preferential attachment on active edges.
(Evolution of I) With probability $1 - \beta$, a new node v is added to I following a symmetrical process, adding c_i edges to v.	**(Edges via evolution of I)** With probability $1 - \beta$: A new edge is added between two nodes active v_1 and v_2 if the new a new common interested has been added

Table 2: Informal description of the evolving model, for a detailed description refer to [14].

8. CONCLUSIONS

We have studied the dynamics of user departure from online social networks and collaboration networks, from the perspectives of local and global network structure. We considered the predictive power of local neighborhoods on the behavior of nodes as well as studied global changes in the network topology. At the local level, we studied individuals and the dynamics in their local neighborhood, measured the probability of user arrival and departure in relation to the activity of their friends. Our findings are threefold: first, there is a strong clustered effect in the timing of departure among friends while this is not as visible in arrivals; second, although both numbers and fractions of neighborhood (in)activity are correlated to the probability of the individual's departure, the fraction of inactive friends has arguably the better predictive power on the departure probability, providing an interesting complement to literature on arrivals which shows number of active friends as the most predictive of these measures; third, once a significant fraction of friends depart, the overall connectivity of individuals in the entire network does not have predictive power as to whether the user will leave the network. At the global level, we looked at the trend of network topological properties over the past few years, showing that as the network evolves, users at the peripheral region of the network are more likely to depart in groups; yet an internal core of the network survives and densifies over time.

We want to emphasize that our results do not prove a causal relationship between the departure of friends and the departure of a user. What we have observed and modeled is only correlated actions among neighbors, and we are aware of the possible factors that can contribute to, or actually lead to, the departure of users. For example, internally, users with similar personal traits may tend to leave the network altogether; externally, the emergence of competing services may draw users away from the original network. Our model provides *one* possible explanation on the emergence of observed pattern, but does not exclude other explanations. Our goal is to offer a space for building better models of how people tune out of social networks, in additional to how they sign up.

9. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD'08.*

[2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 2009.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06.*

[4] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM '11.*

[5] K. Bhawalkar, J. Kleinberg, K. Lewi, T. Roughgarden, and A. Sharma. Preventing unraveling in social networks: The anchored k-core problem. In *Automata, Languages, and Programming*, pages 440–451. 2012.

[6] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *APPROX '00.*

[7] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD '08.*

[8] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *EDBT'08.*

[9] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 2005.

[10] T. Fenner, M. Levene, and G. Loizou. A stochastic model for the evolution of the web allowing link deletion. *ACM Trans. Internet Technol.*, 2006.

[11] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference.* 2002.

[12] M. Jerrum and A. Sinclair. Conductance and the rapid mixing property for markov chains: the approximation of the permanent resolved. In *STOC '88.*

[13] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06.*

[14] S. Lattanzi and D. Sivakumar. Affiliation networks. In *STOC'09.*

[15] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 2007.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08.*

[17] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs

over time: densification laws, shrinking diameters and possible explanations. In *KDD '05*.

[18] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *HYPERTEXT '06*.

[19] M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 2002.

[20] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SDM*.

[21] E. M. Rogers. *Diffusion of Innovations*. Simon and Schuster, 2003.

[22] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion. In *WWW '11*.

10. APPENDIX

THEOREM 2. *There exist small constant values of λ, α and ϵ for which our model densifies in time with high probability.*

PROOF. In order to prove the statement we first analyze the evolution of the high degree interests, in particular we will prove that those interests will continue to grow even after the departure of some nodes from the network. Then we will show that this in turn implies that the network densifies in time.

Let S be the set of nodes in I in $B(P, I)$ such that their degree is bigger or equal to t_d^{δ}, for a fixed constant $\delta = \frac{1}{10}\left(4 + \frac{c_i \beta}{c_p(1-\beta)}\right)^{-1}$. Note that Theorem 1 in [14] implies that $S \neq \emptyset$. Consider a node v in S, note that from construction the $N(v)$ in $B(P, I)$ are a community of size bigger or equal than t_d^{δ} in $G(P, E)$. Let us call this community $C(v)$.

We will start by showing that the number of active nodes in $C(v)$ for an interest $v \in S$ grows asymptotically as in the model without deletions, finally we will use this property to prove the densification.

Let us call $E_{\odot}^{n}(v)$ the expected number of active nodes in $C(v)$ at time n, $E_{\otimes}^{n}(v)$ the number of inactive nodes, e_n the number of edges in $B(P, I)$, and $C_t(v)$ the set of nodes in $N(v)$ as of time t.

We start by bounding $E_{\otimes}^{n}(v)$. We first upper bound the nodes that become inactive by random choice independently of their neighbors. Let us call them $I_{\otimes}(v)$.

Let W_t be the number of nodes (active or inactive) in $C(v)$ at time t. By Lemma 2 in [14] we have that $W_n \leq k|C_{t_\delta}(v)|$ for some constant $k > 0$.

Now we can bound $E[I_{\otimes}(v)] \leq \sum_{t=t_d}^{n} W_t \frac{\alpha}{n} \leq W_n \frac{\alpha}{n}(n - t_d) = (1-\epsilon)\alpha k|C_{t_\delta}(v)|$. In addition, by applying the Chernoff-Hoeffding's bound we have that $I_{\otimes}(v) \leq (1+\gamma)(1-\epsilon)\alpha k|C_{t_\delta}(v)|$ with probability $1 - o(\frac{1}{n^2})$ for any $\gamma > 0$. So by fixing $\alpha < (1+\gamma)\frac{k(1-\epsilon)}{10}$ we have that $I_{\otimes}(v) \leq \frac{|C_{t_\delta}(v)|}{10}$ with high probability.

We now bound the number of nodes that became inactive at time t because other nodes in their neighborhood have left the network, we call those nodes $D_{\otimes}^{t}(v)$. To do it we will first prove that for a node in $C(v)$ the probability of becoming inactive because other nodes in their neighborhood is dominated by a simpler random process and use it to upper bound $D_{\otimes}^{n}(v)$ with $\frac{C_{t_\delta}(v)}{10}$.

First note that until $D_{\otimes}^{t} \leq \frac{|C_{t_\delta}(v)|}{10}$, $D_{\otimes}^{t}(v)$ is dominated the random variable X^t that counts the number of heads that are observed when a biased coin, that gives head with probability $\frac{1}{\left(\frac{8C_{t_\delta}(v)}{10}\right)^{\lambda}} = \frac{1}{\left(\frac{8(\epsilon n)^{\delta}}{10}\right)^{\lambda}}$ when the coin is flipped t times.

Note that by fixing $\lambda \geq \frac{1}{\delta}$, we have that X^t in expectation is ≤ 1 for every $t \leq n$ and thus by the Chernoff bound with probability bigger than $1 - o\left(\frac{1}{n^2}\right)$ $X^n \leq \log n$. But $\log n << \frac{C_{t_\delta}(v)}{10}$, thus $X^n << \frac{C_{t_\delta}(v)}{10}$ with probability bigger than $1 - o\left(\frac{1}{n^2}\right)$.

Now $D_{\otimes}^{t_\delta}(v)$ is equal to 0 and it is dominated by X^{t_δ} until $D_{\otimes}^{t}(v)$ it is smaller than $\frac{C_{t_\delta}(v)}{10}$, but because $X^n << \frac{C_{t_\delta}(v)}{10}$ this implies that $D_{\otimes}^{t}(v)$ is always dominated by X^{t_δ}a. Thus $D_{\otimes}^{n}(v) << \frac{C_{t_\delta}(v)}{10}$ with probability bigger than $1 - o\left(\frac{1}{n^2}\right)$.

So we get:

$$E_{\otimes}^{n}(v) = I_{\otimes}(v) + D_{\otimes}(v) \leq \frac{1|C_{t_\delta}(v)|}{5}$$

with probability at least $1 - o\left(\frac{1}{n^2}\right)$. Hence by union bound and by the fact that $|S|$ is smaller than n we get that with high probability for all the $v \in S$ $E_{\otimes}^{n}(v) \geq \frac{1|C_{t_\delta}(v)|}{5}$.

Now we want to study how $E_{\odot}^{n}(v)$ evolves in time. To do it, note that by the bound of $E_{\otimes}^{n}$ we have that the number of deleted node in $C_t(v)$ is $\leq \frac{1|C(v)|}{5}$ for all $t \geq t_\delta$. Thus the grow of $C(v)$ will dominate the grow of a community that has size $\frac{4|C(v)|}{5}$ at time ϵn in the process without deletion of the nodes. By combining the above property and Lemma 2 in [14] we get that for any $C(v)$ with $v \in S$ the final size of the cluster in the process without deletion and in the process with deletion are just a constant factor away.

Now note that from [14] we know that in the process without deletion the graph densifies because

$$\begin{aligned}
|E_n| &> \sum_{v \in I} (\text{edges generated by a cluster of size I}) \\
&> \sum_{v \in V} (\text{edges generated by a cluster of size I}) \\
&> \sum_{v \in S} (1-\sigma)p(|C(v)|(|C(v)| - 1)) \\
&> \sum_{i=n^{\delta}}^{n} (\#\text{ of nodes of degree } i \text{ in } I)(1-\sigma)pi^2) \\
&> \sum_{i=n^{\delta}}^{n} \left(\left(\left(\frac{n}{\zeta(-2-\Delta)} \frac{1}{(i)^{2+\Delta}}\right)(1 \pm o(1))\right)(1-\sigma)pi^2\right) \\
&\in \omega(n)
\end{aligned}$$

Where σ is a constant >0, $\zeta()$ is Riemann zeta function and $\Delta = \left(4 + \frac{c_i \beta}{c_p(1-\beta)}\right)^{-1}$.

Thus the densification property in the process without deletion is a consequence of the number of edges in the clusters generated by large interest but we have proved that the size of those interest change by at most a constant factor, so also in the model with deletion the number of the edges is $\omega(n)$ and thus the model densifies in time. □