Influence of First Steps in a Community on Ego-Network: Growth, Diversity, and Engagement

Atef Chaudhury[*]
University of Waterloo
Waterloo, ON N2L 3G1,
Canada
a29chaud@uwaterloo.ca

Myunghwan Kim
LinkedIn Corporation
Mountain View, CA 94043,
USA
mukim@linkedin.com

Mitul Tiwari
LinkedIn Corporation
Mountain View, CA 94043,
USA
mtiwari@linkedin.com

ABSTRACT

Social networks are important in our offline and online life. In particular, networks that people form within certain communities can be critical for their engagement and growth in those communities. In this work, we analyze the growth of ego-networks for new employees of companies in LinkedIn, and study how the pattern of network formation in a company affects one's growth and engagement in the company.

We first identify some key common patterns – such as functional group preference and triad-closing propagation – in the network formation within the company. We find that because of these common patterns initial connections that people make in new organizations become crucial. To be specific, our analysis demonstrates that first few connections in a newly joined company have great impact on characterizing future status in the company, which includes the total number of connections, network diversity, and retention.

Keywords

social network, network growth, social engagement, organizational behavior

1. INTRODUCTION

Understanding the evolution of an individual's network in a new community is important to help grow social network of the individual. This understanding is valuable for figuring out missing edges and link prediction [3,15,17,24] in any online social network since any online social network is partially observed, that is, two people may know each other but may not be connected with each other on the online social network. Hence many online social networks recommend other connections in order to grow network of an individual. For example, LinkedIn, the largest professional network, exposes its connection recommendation system through "People You May Know" feature.

Understanding growth of ego-network, connections of an individual and network of connections between them, and engagement in a recently joined community is another important challenge. For example, how does a new employee in a company network or connect with other employees of the company? Are there patterns in initial connections in the new company that influence future retention in the company? Is there any homophily in the company network of a new employee? These are important questions for any social community, not just a company network, such as online groups and online school networks on a social network.

Despite the importance of the above questions, there is a little understanding on how the ego-network of a new member of a community evolves within the community. In this work we aim to address those questions by analyzing LinkedIn's company network data, which consists of more than 360 million members and millions of companies. To our best knowledge, our work is the largest analysis of individuals' networking behavior in a company, and LinkedIn's data is ideal to perform such a study to understand the ego-network growth of a new member in a company.

In the first line of investigation, we find that a significant fraction of an individual's connections in a company form through closing triangles from their first few connections. This phenomenon can be modeled as propagating connections through triangle closing. For instance, when someone joins a company, his or her manager introduces the new employee to other employees. Afterwards the new employee can expand connections through their existing connections but not directly through the manager. We further analyze this propagation of triangles in a company network and find that a large number of connections for an individual are formed by such triangle closing propagation.

In the second line of analysis, we find that both triangle closing and its propagation makes first connections crucial for making a new employees grow and engage in a new company. Such first connections can help the new employee reach more people in the company. Also, if established well by the help of those first connections, then the new employee is able to grow quickly in the company. Therefore we can imagine that first few connections have great impact on life in the new company. Our analysis reveals that first few connections influence a new employee's later status in three aspects: network size, network diversity, and retention.

Network Size: we find that a new employee's network size is affected by the number of connections that the first few connections of the new employee have. Our analysis shows that triangle-closing and its propagation is a primary

[*]This work was done when he worked for LinkedIn.

WWW '16 Montreal, Quebec Canada
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

vehicle for a new employee to expand the network. Therefore, if the first few connections are already connected to enough number of people, then the new employee has more opportunities to add connections through triangle-closing or its propagation. On the other hand, if the first few connections are not connected with enough people in the company, then ego-network growth of the new employee through triangle-closing is limited, and as a result the new employee is likely to have a smaller number of connections later.

Network Diversity: we discover that we can predict diversity of a new employee's ego-network with respect to functional groups, which represent the roles of jobs in companies (such as engineering, product, and sales) by looking at the first few connections. For instance, if the first few connections of a new employee belong to diverse functional groups, then the future ego-network of the employee is likely to remain diverse in terms of functional groups. This phenomenon is not consistent with homophily [26], which implies that people tend to be connected with those in the same functional group. Our analysis demonstrates that homophily is a collective behavior of people while each individual does not necessarily form connections through homophily, which is an interesting insight.

Employee Retention: we find that the first few connections also influence the retention of a new employee in the company. Our analysis shows that a new employee having high-degree and more senior people in the first few connections is less likely to leave the company in 1.5 years. Note that the retention is not directly related to any networking behavior but correlated with the composition of the first few connections. This analysis agrees with our intuition that the establishment of good social network in a newly joining company is important for the professional life in that company.

The rest of paper is organized as follows. Section 2 reviews previous study of ego-network analysis and company-network analysis. Section 3 describes the connection pattern in a social network, and Section 4 follows with narrowing down the focus to the ego-network growth in detail. Finally, Section 5 discusses our main contributions about the influence of first connections on network size, network diversity, and employee retention.

2. RELATED WORK

As social networks are regarded critical in many places, research about the networking behavior of an individual has been extensively performed. Triangle-closing [20, 22, 30], homophily [19, 26, 32], structural diversity [16, 29], and other network topological structure [2, 14] have been studied as main mechanisms in the formation of connections. Such mechanisms have been used in many applications, including recommendation of new connections [3], prediction of network properties [17], and inference of user properties [18]. In this work, we extend triangle-closing to its propagation, which means that one triangle-closing opens up an opportunity of another triangle-closing that would not happen without the former triangle-closing. We also show that homophily can behave differently if it is combined with triangle-closing. Given the extended analysis on the network formation mechanisms, we demonstrate that first few connections of a new employee in a company influence his or her later status – such as network size and engagement.

On importance of social networks, previous studies have revealed that social networks in organizations influence an individual's performance [4, 27, 31], personal regard [5], and turnover [10]. In particular, structural diversity [4, 6, 27], culture [28], relationships with supervisor [8], and strong ties [13, 25] have been shown as key factors for better performance and engagement in the organizations. However, most of this previous study has been performed by surveying subjects in certain organizations; hence, the number and the diversity of subjects is inevitably limited. To our best knowledge, our work on LinkedIn dataset is the largest analysis that uses the most number of subjects and organizations. Moreover our work envisions the predictive power of initial connections in an organization, while the previous work typically shows the correlation between a certain characteristic of a person – such as performance – and some measurements like structural diversity [4] in the final network.

3. CONNECTION PATTERNS

Social networks in the real-world are known to be formed by some common patterns. In particular, *triangle-closing* [20, 22, 30] and *homophily* [15, 26] are well-studied and key mechanisms in the formation of network connections. Triangle-closing indicates that employees are likely to connect to their second-degree connections (*i.e.,* friends-of-friends). Homophily implies that employees tend to form connections with others who are similar to themselves – such as being in the same organization and having the same hobby.

In this section, we use within-company networks of large technology companies in LinkedIn dataset, and aim to investigate the patterns of forming network connections within companies. We introduce the concept of propagation in the triangle-closing and examine structural patterns in such propagations. Also, while showing homophily on functional groups in companies, we combine homophily and triangle-closing to see the joint patterns between the two phenomena.

3.1 Triangle Closing and its Propagation

Triangle-closing, the formation of new connections with one's second degree network, plays a strong role in determining the structure of social networks. Furthermore the closing of one triangle can lead to the creation of more potential triangles, which may then also lead to the creation of other potential triangles. This can be viewed as the propagation of connection through triangle-closing, as the creation of a connection closing one triangle influences the creation of another creation closing another triangle. This phenomenon would then have the potential to create cascades, similar to those of information or invitation diffusion [1, 12, 23].

To study this propagation thoroughly, we define a new representation of a network, named an *E-Graph.* In an E-Graph, each node corresponds to each single connection in the original network. To avoid confusion, we denote the node in the E-Graph by *E-Node.* A directed edge in the E-Graph, called an *E-Edge*, is then formed when a source and a destination E-Nodes (*i.e.,* connections in the original network) are the second and the third connections of a certain triangle in the original network, respectively. In this way there is a one-to-one relationship between E-Edges in the E-Graph and triangles in its original network.

Figure 1 illustrates the E-Graph representation. On the left side, given the existing connections between B∼E (black), the order of connections A-E (red), A-C (green), and A-B (blue) is shown in the real network where the number

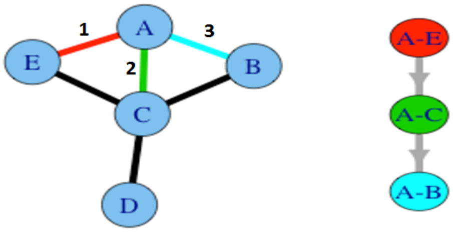

Figure 1: An example of E-Graph representation. The network on the left shows the formation of three connections A-E, A-C, and A-B in this order, given the existing connections among B∼E. This ordering translates to the E-Graph on the right as A-E creates the triangle A-E-C which is closed by A-C. As A-C is propagated by A-E, it is a child of A-E in the E-Graph. Similarly A-B is a child of A-C as A-B.

marked on each edge indicates the order among these three connections. On the right side, we draw the partial E-Graph representation of the left network only between these three connections. As each connection of A-E and A-C is respectively the second and the third connection of the triangle A-E-C, an E-Edge is drawn from E-Node A-E to E-Node A-C. Similarly, A-C and A-B are the second and the third connection of a triangle A-C-B, so we connect from A-C to A-B in the E-Graph. The reason for attributing the second connection of a triangle to the parent of the corresponding third connection is that the third connection would not be likely to be formed without the existence of the second one.

As an E-Graph captures the propagation pattern of triangle-closing in the original network, the structural characteristics of the E-Graph can provide insights into the role which triangle-closing plays in forming connections, similarly to information or invitation cascade analysis [1, 7, 11, 12]. Since many metrics in the cascade analysis – such as depth – are defined over trees, we sample E-Nodes at random and investigate the structure of sub-components reachable by the sampled E-Nodes. Note that the sub-component of a certain E-Node does not allow any ancestors of the given E-Node, while containing all the descendants. The structure of each sub-component can be then quantified by some cascade metrics – such as maximum depth, and Wiener Index [1, 11].

For comparison, we use the following model as a baseline. First we generate a random network by shuffling the timestamps associated with each connection in a real network. This way preserves all the connection patterns in the final state but changes the order of connection creation and the triangle-closing patterns. We then obtain the E-Graph for this random network. The different triangle-closing patterns in the random network will result in the different structural patterns in its E-Graph. Hence, by comparing the structural patterns between the E-Graphs of real and random networks, we can find unique patterns in triangle-closing and its propagation for the real network.

Here we use three metrics to quantify the pattern of triangle-closing propagation in each sub-component. First, we define the size of a sub-component by the number of E-Nodes in it. Second, we measure the maximum depth by finding the maximum path length in each sub-component. While the size implies the amount of triangles that a connection can propagate, the maximum depth indicates the length of such propagations. Last, we also measure the Wiener Index, which is computed by the average path length between two E-Nodes of a sub-component. Wiener Index represents the structural

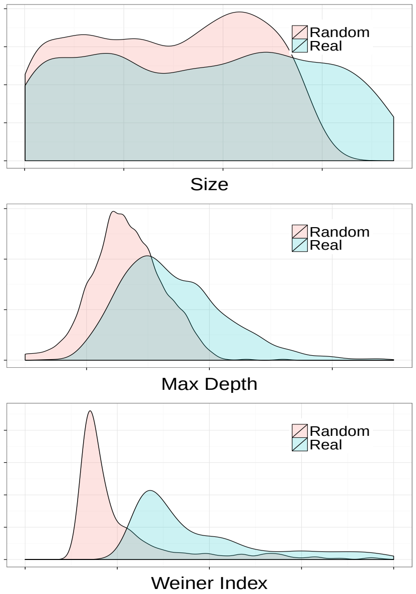

Figure 2: Structural patterns in the E-Graph of company networks. From sub-components of sampled E-Nodes, we plot the probability density of size (top), maximum depth (middle), and Wiener Index (bottom). We compare each pdf in the real E-Graph (blue) with the corresponding pdf in the random E-Graph (red). Each plot implies that the real network is more viral in the triangle-closing propagation than the random network.

virality [1], the value of which is high for chain-like structure but low for broadcasting-like structure.

Figure 2 shows probability density plots with respect to these three metrics for sub-components of E-Nodes in an E-Graph. Each blue area represents the density from the E-Graphs of company networks in LinkedIn, whereas the corresponding red area plots the density from the E-Graphs of their random networks. In every plot we observe that the density for the real networks is more skewed to the right than the density for the random networks, *i.e.,* each metric for the real networks tends to be high compared to the random networks. The distinction between real and random networks is the most visible in the Wiener Index. Our observations demonstrate that the deep propagation of triangle-closing appears in the real-networks whereas it does not necessarily happen under the random situation. Therefore, in the real-world there exists the phenomenon that the closing of one triangle leads to the closing of other triangles.

Our analysis on the propagation is important for two reasons. First, our results demonstrate that triangle-closing plays a role not only in forming one connection but also in leading to more connections later through other triangle-closing. Second, because of this propagation, initial connections become very critical for the growth of a network. In particular, such initial connections for a person can be crucial for the growth of his or her ego-network, and we will study the influence of triangle-closing propagation on ego-networks in Section 4.

3.2 Homophily

For another connecting behavior, homophily is a phenomenon that people of the same characteristics – such as gender or socio-economic background – are likely to be connected [26]. Within company networks, each employee is likely to con-

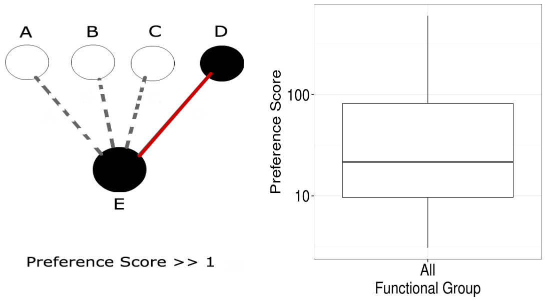

Figure 3: Functional group preference. Left figure illustrates an example to show that only D is connected so the preference score of the black functional group is 4. The right plot represents a box plot of such preference scores for all the functional groups in company networks in LinkedIn. We see significantly high preference scores, *i.e.,* a strong homophily in the functional group.

nect with others doing the same kind of jobs. For instance, engineers are likely to be connected with other engineers as they are likely to collaborate. Hence, homophily can be observed with respect to functional groups, which represent the roles or categories of jobs in companies. Some examples of functional groups are engineering, marketing and HR.

To verify homophily on the functional group, we use the same company networks in LinkedIn as the previous analysis. We define a *preference score* of each functional group for the notion of homophily as follows. For a functional group F, we compute the probability that an employee is connected to others within the same functional group. We then divide this probability by the portion of the given functional group among all and refer to this value to the preference score:

$$\text{Preference}(F) = \frac{P\left[(x, y) \in E | x, y \in F\right]}{(|F| - 1)/(|N| - 1)}$$

for a set of nodes N and edges E in the given network. The denominator part indicates the connection probability within a functional group under the situation without homophily. In short, the preference score normalizes the connection probability within a functional group by the size of that functional group. For example, in the left plot of Figure 3, suppose that only D and E are connected where each of black and white circle corresponds to a different functional group. To compute the preference score of the black functional group, we first have $P(\text{black}) = 1$ for the connection probability within the functional group. Now we need to normalize this probability by the size of functional group. Without homophily, the probability of connecting to D would be $\frac{1}{4}$ for the given black node E. Hence, by definition, the preference score is $\text{Preference}(\text{black}) = 1/\frac{1}{4} = 4$.

We obtain those preference scores for all functional groups in the LinkedIn dataset and illsturate their distribution as a box-plot on the right of Figure 3. Note that the y-axis is presented on the log-scale. While a certain functional group shows very strong homophily, we see that all functional groups have preference scores higher than 1. Therefore, this result demonstrates that functional group homophily plays a strong role in forming the connections within a company network. This result is also intuitive with respect to company networks, as employees are typically organized in teams which share a common purpose and thus are likely to connect with employees in their own functional groups.

3.3 Homophily and Triangle Closing

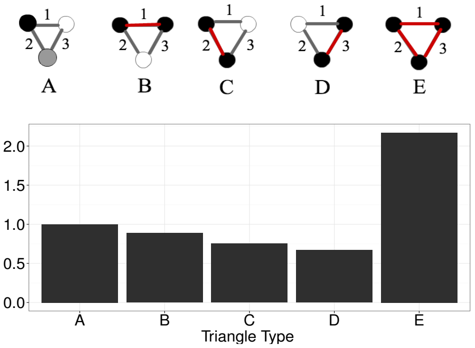

Figure 4: Proportions of five types of triangles. The x-axis values of the bar graph correspond to each of the triangle categories illustrated in the top diagram. The relative number of triangles in each triangle type is shown on the y-axis. Type E in which all connections is the most common as a strong homophily. However, note that Type B is 20% more common than Type D. If the first connection is between two different functional groups, then triangle-closing is less likely to happen.

So far we have investigated triangle-closing and functional group homophily. However, these two phenomena can jointly happen in the network formation. To examine this joint effect, we categorize triangles into 5 types considering functional groups and the order of connection formation as illustrated on the top of Figure 4. Each diagram represents the functional group by color and the order of connections by a number next to each connection. Type A indicates a triangle among all different functional groups, where Type E means a triangle within the same functional group. Note the difference between Type B, C, and D. All of them represent the same composition that two connections in a triangle comes from the same functional group while the other is from a different one. The only difference is the order of connections in a triangle. As there is only one connection from the same functional group in this composition, Type B, C, and D are distinguished by whether the same functional group connection is formed first, second, or third.

The bottom bar plot in Figure 4 draws the relative proportion of each of these 5 types. The proportions are normalized by dividing by the number of triangles in Type A (*i.e.,* Type A's value is 1). Type E, a triangle within the same functional group, leads to the largest proportion of the total triangles with over twice as many as triangles in Type A. The large difference in the relative numbers between Type A and E affirms functional group homophily as previously discussed. As employees within the same functional group tend to connect with one another, there is also a higher likelihood of creating triangles within the same functional group than across functional groups.

We also note that the proportions of Type B, C, and D are statistically significantly different. For instance, Type B where the same functional group connection occurs first exists 20% more than Type D where the same functional group connection comes last. The differences between Type B, C and D illustrate a case in which functional group homophily is hampered. The comparison between Type B and D shows that connecting to an employee of the same functional group through an employee of a different functional group is less likely than connecting to employee of a different functional group through employee from the same one. This result implies that an employee is likely to continue to connect to those outside their own group by connecting to employees of different functional groups. Under a company setting, those who initially connect outside of their functional group may be better able to leverage these connections to continue to connect outside of their functional group.

4. EGO NETWORKS

So far we have characterized the primary mechanisms of network formation within companies. Such mechanisms can have great effects on how each employee expands his or her network within a company.

In this section, we draw on those findings and focus on how they manifest within an employee *ego-network* in each company. To study these ego-networks, we select top 500 companies in LinkedIn by their average degrees, and investigate the pattern of each employee's ego-network growth within each company. For convenience, we refer to one's ego-network within a company as *company-centric ego-network*.

Our dataset has three key properties. First, a company organization can be regarded as one of the most explicitly defined communities. From the graph theory perspective, the density of connections within each company is likely to be larger than the density across companies [15]. Thus, some findings in these company networks may be projected to broader communities. Second, LinkedIn users from those top 500 companies tend to be highly engaged with LinkedIn so that they reflect their real-world company-centric ego-network into LinkedIn connections well. Hence, the analysis on our dataset is beyond online world and can illustrate real-world human behavior to large extent. Third, the number of employees in the top 500 companies is still over a million and even the new employees in 2013 is in the order of 100k. This is on a much larger scale than survey-based studies [4, 10, 27, 28]. To our best knowledge, this study is the largest analysis with respect to company networks. Using the LinkedIn dataset, we analyze the growth patterns of company-centric ego-networks, particularly focusing on triangle-closing and its propagation.

4.1 Propagation by First Connections

In the previous section, we find that triangle-closing has great importance with regards to the formation of connections and exhibits propagation-like behavior in company networks. Our observation on the triangle-closing can be leveraged to gain insights about smaller scale ego-network growth within a company. For example, if triangle-closing is one of the primary vehicles by which employees make connections, and if it enables employees to close more triangles, then we can imagine that an employee's earlier connections will have a strong influence on their later connections.

To investigate the influence of an employee's first few connections on the formation of later connections, here we focus on triangle-closing through the first few connections in an employee's company-centric ego-network. This investigation will give us an idea about whether the growth of each individual employee in a company is independent of the others or influenced by the others.

We quantify this influence from first connections by what proportion of an employee's ego-network after two years is propagated from them. In this context, "propagated" refers only to connections made with the employee's first connections' existing connections, *i.e.,* through triangle-closing. Note that the propagation does not necessarily capture all the common connections between the employee and the first connections. If another employee is connected with the given employee prior to the first connections, then this common connection is not counted as propagation.

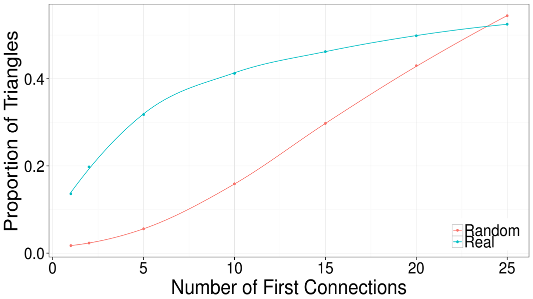

Figure 5: Coverage by triangle-closing through first connections. The number of first connections is on the x-axis and the proportion of triangles covered by those first connections is on the y-axis. The blue curve depicts the tendency in the real network, whereas the red line represents the random network results in which the connection order is shuffled. Note that first few connections play a more important role in triangle-closing in the real network than in the random network.

Under this definition, the blue curve in Figure 5 plots the percentage of propagation from first connections to an employee's network after 2 years, as a function of the number of first connections. We observe that approximately 40 percent of an employee's final network is propagated from their first 10 connections within the company.

However, the blue curve itself cannot imply whether the 40 percent of connections are propagated just because the first connections have some homophily such as the same functional group or because the first connections have real influence. To distinguish the influence from the homophily effect, we perform the same analysis on the random network introduced in the previous section, which shuffles the ordering of connections in the original network. If most of this propagation is driven by homophily, the change of connection order should not affect the level of propagation.

The red curve in Figure 5 represents the proportion of first connections' propagation in the random network. We note that there is a large gap between the real network (blue) and the random network (red) with respect to the first connections' propagation. For example, less than 10 percent of final connections are propagated by first 10 connections in the random network, while the real network shows about 40 percent for the same measurement. This supports the hypothesis that an employee's first connections have a strong influence on his or her final company-centric ego-network. Hence the connection preferences of an employee's first connections affect the employee's connection preferences.

In the real-world, when people newly join some communities, they may start by making a few initial connections and gradually grow their networks through existing connections. Particularly in a company, new employees meet their team members first and they can be introduced to some other employees in formal meetings or informal gatherings.

4.2 Influential Attributes on Propagation

Given the significant propagation of first connections on the future ego-network, now we examine the attributes that drive this propagation. We thus aim to identify the situation where the propagation through triangle-closing affects the formation of connections the most for an employee.

In order to explore such attributes, we again bring up the concept of E-Graph and E-Node in Section 3.1. Recall that E-Graph is the representation of triangle-closing and its propagation. E-Node and E-Edge is a connection and a triangle-closing in the real network, respectively. By definition, the out-degree of an E-Node in a E-Graph refers to the number of connections that a connection corresponding to the E-Node propagates. Hence, if we can find the relationships between some attribute of an E-Node and its out degree, the found attribute can be a key for the propagation through triangle-closing.

Here we focus on the following two attributes of a connection, *i.e.,* an E-Node: the degree sum and difference of two connectors in a connection (E-Node). Note that the combination of these two attributes contain enough information to determine the degrees of two connectors. For instance, the high degree sum and low degree difference means that both connectors have high degrees.

Regarding measurement of triangle-closing propagation, the out-degree of an E-Node may not be good for capturing the contribution of the given E-Node to triangle-closing. For instance, suppose that an E-Node represents a connection between employees A and B, and a following connection between A and another employee C closes a triangle A-B-C. According to the E-Graph representation, A-C is the child of A-B. Under this situation, the triangle A-B-C gives full credit to the out-degree of A-B regardless of the number of common connections between A and C. If A and C had a lot of common connections before their connection was formed, then it is hard to say that the connection A-B necessarily influences the triangle-closing connection A-C.

To tackle the above issue, we use the normalized out-degree of an E-Node x in the following way.

$$\text{NormOutDeg}(x) = \sum_{x \to y} \frac{1}{\text{InDeg}(y)}$$

where each E-Edge $x \to y$. This normalized out-degree basically gives the even credit to each potential E-Node for a given child E-Node. In other words, all the second connections of triangles that the given connection closes uniformly share the contribution for the corresponding triangle-closing. Therefore, high normalized out-degree of an E-Node implies that the corresponding connection actually leads to many triangle-closings.

To find the relationships between the attributes of an E-Node (degree sum and difference) and the normalized out-degree, we first put each E-Node into 5 × 5 bins based on the values of degree sum and degree difference. We then compute the average normalized out-degree of all the E-Nodes in each bin. For comparison, we also use the random network that shuffles the ordering of all the connections, which preserves the triangle patterns but does not maintain the pattern in the E-Graph. If a certain bin in the real network shows visibly higher average normalized out-degree than in the random network, then this higher average indicates that the connections (E-Nodes) falling in this bin tends to draw more triangle-closing than expected.

Figure 6 shows the ratio of the average normalized out-degree between the real network and the random one. This heatmap is divided by 5-by-5 bins, each of which represents the same level of degree sum (column) and difference (row) of two connectors. For a fixed row, there is a decreasing trend of the relative value as the degree sum increases. On the other hand, for a given column, we also observe an increasing ratio value as the degree difference increase. Note that the existence of low degree connector results in low degree-sum for the same degree-difference and high degree-difference for the same degree-sum. In Figure 6, the leftmost bin in each row and the top bin in each column falls into the case where one of two connectors has low degree. Therefore a connection involving low degree connector is expect to lead more triangle-closing.

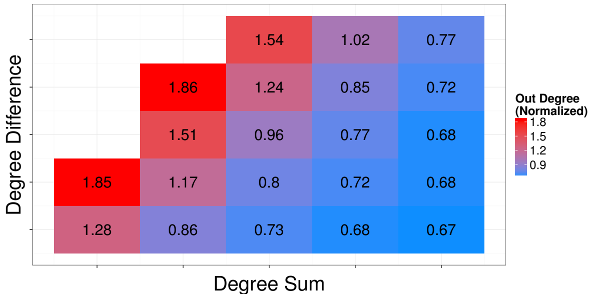

Figure 6: Influential triangle-closing. For each connection, we compute its contribution to the triangle-closing by the normalized out-degree of its corresponding E-Node in the E-Graph. Such contributions are measured for both real and random networks, and then this color-map visualizes the relative value of contributions in the real network to the random network, segmented by degree sum (x-axis) and degree difference (y-axis) of two connectors. The red area along with diagonal axis shows the importance of connections between low and high degree connectors, and thus implies the importance of first connections to new employees.

In overall, E-Nodes show higher normalized out-degree values when the degree-difference is high, and degree-sum is low. For these conditions to be met, an E-Node needs to consist of one connector with very low degree and the other connector with very high degree. This would indicate that connections with high degree employees propagates a high number of connections in the future, especially if the other connector is of low-degree.

From these phenomena, we have the following hypothesis. As the degree of a new employee in a company is likely to be low, the propagation through triangle-closing is more critical for the new employee to form connections within the company. Similarly, we might guess that high degree connections are capable of more propagations. Thus whom a new employee is connected with in the beginning can be influential on the ego-network of the new employee later.

Anecdotally, within a company network, this situation will typically occur when a new employee (low degree) joins a team and a manager (high degree) introduces the new employee to the other employees. This case is an example of a connection truly leading to triangle-closing. In this case, if the manager is well-established within a company so that (s)he can introduce more employees in the future, then the new employee can form more connections through triangle-closing led by the connection with the manager. On the other hand, if the manager is also relatively new to the company, then the manager may not be able to introduce many other employees; that is, the connection with the manager (low degree-sum, low degree-difference) cannot lead to lots of triangle-closing.

Our analysis in this section demonstrates that the propa-

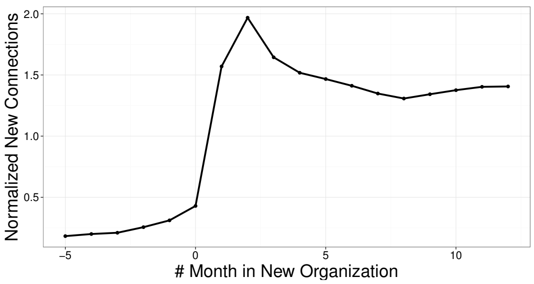

Figure 7: New connections around job change. While x-axis points to the number of month since a new employee joins, y-axis indicates the relative number of new connections in each month to the overall average number of new connections per month. A spike is seen around the joining time and the growth is quickly stabilized.

gation through triangle-closing plays a more important role in forming later connections when a low degree employee is involved. This analysis inspires the hypothesis about a new employee, which provides the motivation to the next section. In the next section, we will study what the first connections can influence and view the first connections as a predictor of future company-centric ego-network.

5. INFLUENCE OF FIRST CONNECTIONS

A hypothesis we have in the previous section is that the first connections of a new employee influence the formation of later connections. If our hypothesis is true, then the first connections can affect some characteristics of the new employee in the future. The characteristics may be related to the later company-centric ego-network of the new employee, or may be network-independent, such as employee retention.

To verify our hypothesis, we aim to find predictive information in the first connections of a new employee with regards to the future characteristics of the new employee . Here we focus on three kinds of characteristics: network size, network diversity, and employee retention. While the first two characteristics are directly related to network formation, the retention is apparently unrelated to the network formation. Our following analysis reveals that all of these three characteristics are influenced by the first connections that new employees make when they join new companies.

5.1 Setup

For our analysis in this section, we use the top 500 companies LinkedIn dataset, but choose new employees who joined these 500 companies in 2013. We observe the first 10 connections of those new employees and find the relationships between these 10 connections and the characteristics of the new employees after 1.5 years. We use 10 as the number of first connections, as Figure 5 shows that a high proportion of the total company-centric ego-network overlaps with first 10 connections' company-centric ego-network as well as the first 10 connections play more significant roles in the real network than the random network. To make first 10 connections be significantly less than final number of connections, all the employees who end up with less than 20 connections after 1.5 years are removed in our analysis.

One benefit to using first connections as a predictive measure, is that they do not take significant time to form after the employee joins the new company. Figure 7 shows that there exists a spike in new connections during the first three months after joining the new company. This spike is likely due to the burst of new people an employee meets shortly after joining a new company. Due to this influx of new connections, a small time frame is required before predictive value can be gained from an employee's first connections.

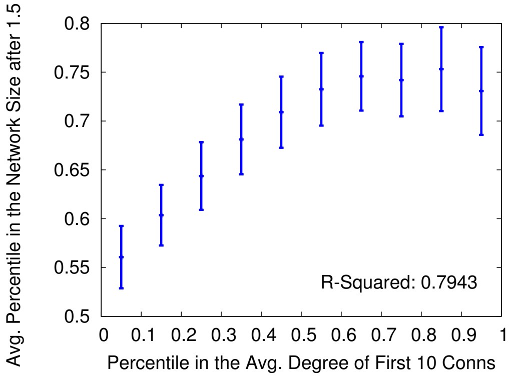

Figure 8: Influence of first connections on network size after 1.5 years. The plot depicts the percentile of average degree of an individual's first 10 connections on the x-axis and the average percentile of an individual's network size after 1.5 years on the y-axis. The positive correlation implies that more degrees of first 10 connections influence the final network size.

5.2 Ego-Network Size

As Section 4.2 describes, we expect that first connections are crucial for determining the number of potential triangles an employee will be able to close. If an employee connects to another well-connected individual, then they have created more potential triangles and hence have a greater potential for triangle-closing. This finding gives us the insight that the degree of an employee's first connections can be critical to growing the size of their company-centric ego-network.

In order to verify this insight, for each new employee in 2013, we measure the average degree of a new employee's first 10 connections and compare it to their own degree 1.5 years after joining. The degree of each connection is measured at the time of being connected.

As companies have different propensities for LinkedIn engagement and hence a high range in the number of intra-company connections, we need to normalize the degree in some way. Also, some skewness can be caused by the heavy-tailed degree distribution [9]. To tackle these two issues, we borrow the idea from Spearman's rank correlation [21]; that is, we use the rank (percentile) of each degree value rather than its absolute value. Figure 8 depicts the positive Spearman's correlation between first 10 connections' average degree and network size after 1.5 years. We see a strong positive correlation between the average degree of first 10 connections and network size after 1.5 years. This positive correlation is consistent with our insight that more opportunities for triangle-closing can result in more connections. However, while the correlation is very strong upto a certain level, such first 10 connections' degree effect is saturated above that level. We believe that triangle-closing opportunities above a certain threshold level may not add more connections due to limited time and energy of employees.

5.3 Ego-Network Diversity

The previous analysis with respect to network size verifies that triangle-closing is a primary vehicle in company-

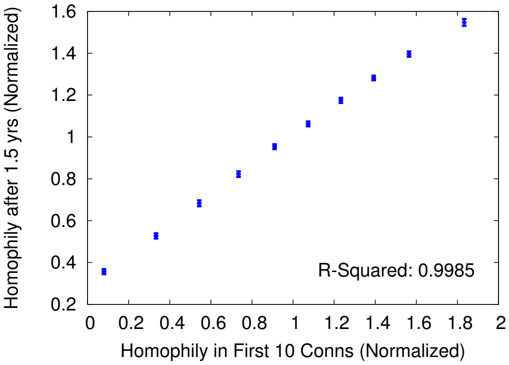

Figure 9: Consistency in functional group homophily. There is a nearly perfect positive correlation between functional group similarity with first ten connections (x-axis) and functional group similarity with all the connections after 1.5 years (y-axis).

centric ego-network expansion and that an employee's first 10 connections are important seed connections in their future company-centric ego-network. We now extend this line of reasoning to examine the effects of homophily, in the form of functional group preference, within company networks. If the company-centric ego-network is formed through the first 10 connections, then the diversity of the first 10 connections is expected to influence the diversity in the later company-centric ego-network. We study this effect as follows.

First, as a notion of homophily, we measure functional group similarity (1 if shares; 0 otherwise) for every connection made by new employees. If functional group information is missing, then we do not use such users in the analysis. The average of the similarity is measured with first 10 connections and then compared with all the employee's total network after 1.5 years. Since each company has a different composition of functional groups, each functional group similarity score is normalized as the relative value to the average value in each company. That is, per company, each similarity value with first 10 connections and with all connections after 1.5 years is divided by its company-wide average value.

Figure 9 shows almost a perfect positive correlation between first 10 connections and all the connections after 1.5 years with respect to functional group homophily. This plot can be understood in two ways. First, as interactions with other functional groups may be consistent over time, the level of functional group homophily may not change a lot for 1.5 years. For instance, as some functional roles like HR are required to make connections with the other functional groups, they continuously form connections with various functional groups. On the other hand, engineers might be mostly connected with engineers since joining until leaving companies. Second, one of our findings in Section 3 shows that a connection between different functional groups drives the network formation opposite to the functional group homophily. Hence, for a new employee, first connections from different functional groups do not lead to triangle-closing in the same functional group as the given employee. This behavior boosts the network expansion in different functional groups while limiting the network expansion in the same functional group, so that it contributes maintaining the level of homophily to some extent.

This result however, can be biased by the fact that the functional group homophily does not account for differences in relative population proportions among the functional groups. Members of functional groups with high population proportions would have greater opportunity to connect within their own functional group, while those in smaller functional groups would be forced to connect to those outside their own functional group.

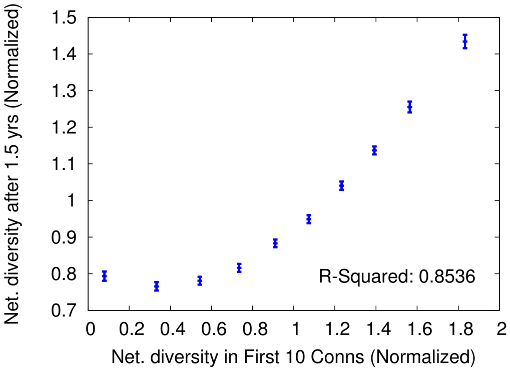

Figure 10: Network diversity. Normalized network-diversity of an individual's first 10 connections (x-axis) is correlated with that of the employee's network after 1.5 years (y-axis). This demonstrates that homophily is a collective behavior of a group of people and may not hold for each individual.

To overcome this bias, we define the network diversity by the probability that a randomly selected pair from the company-centric ego-network do not share a functional group. This measure does not consider the functional group of the employee and hence would not be biased by the relative population proportion of the employee's functional group.

Again, since the network diversity value can vary company by company, we use the normalized value by dividing measures by their corresponding company-wide average value. Figure 10 shows that there exists a positive correlation between the network diversity of the first 10 connections and the network diversity after 1.5 years. High R^2 statistic (0.8536) verifies that the positive correlation is strong.

A similar explanation to the previous analysis can be applied. If an employee contains low network diversity (high homophily) in the first 10 connections, then by a combination of triangle-closing and functional group homophily, the later connections continue to display low diversity. An employee primarily connects to the first 10 connections' connections, who are likely to share the same functional group as the first 10 connections. An employee will thus continue to connect to employees in the same functional group as their first 10 connections. For this reason low network diversity would last in the final ego-network of the employee.

Another explanation would be that an employee's first connections are a sample of that employee's connection preferences. If a employee prefers to connect to others from different functional groups, then that characteristic can be determined by looking at their first 10 connections.

This phenomenon is not consistent with homophily [26], which implies that people tend to be connected with those in the same functional group. Our analysis here is not well in align with homophily [26], because we would expect low network diversity regardless of the diversity in first 10 connections if the homophily governed the primary connecting behavior. We note that homophily is a collective behavior of people though certain individual's may display non-homophilic behavior.

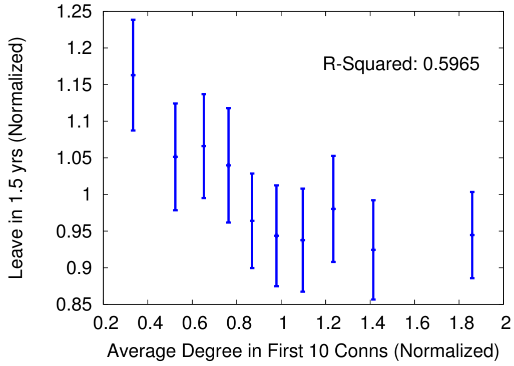

Figure 11: Influence of first 10 connections' degrees on retention. Normalized average degree of an individual's first 10 connections (x-axis) shows a negative correlation with the turnover rate of the individual in 1.5 years (y-axis). Given that first 10 connections' average degree affects the size of social network in Figure 8, this result implies that forming enough social connections is important in the retention.

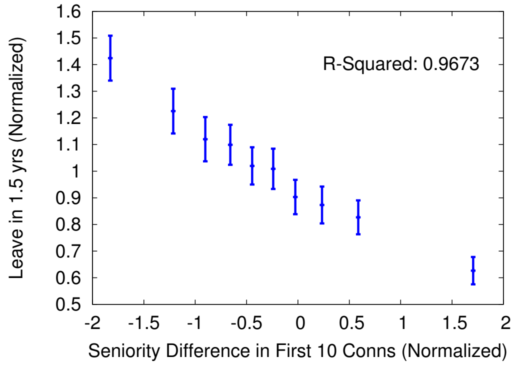

Figure 12: Influence of seniority difference in first 10 connections on employee retention. While the x-axis shows normalized seniority difference between an individual and first ten connections, the y-axis shows the normalized turnover rate of the employee in 1.5 years. The result shows a strong negative correlation indicating that employees whose first 10 connections are of a high seniority level are less likely to leave the company.

5.4 Employee Retention

Apart from predicting the characteristics of an employee's final ego-network, one's first 10 connections can also provide information about an employee's future within the company. For example, if we assume that social networking within a company is important for staying in the company, we can study the relationship between an new employee's first 10 connections and the employee's retention .

Regarding the measurement for the retention, here we use the turnover rate of new employees in 1.5 years, which represents the probability that the new employees leave the company in 1.5 years. Our analysis excludes all the members if company serving time is missing. Note that all the measurements are normalized in the same way as before, as each company has a different level of turnover.

Figure 11 illustrates the influence of average degree of one's first 10 connections on the turnover rate of the corresponding new employee. While variance is high, we can find the slightly negative influence on the turnover rate. This implies a positive correlation between the average degree in first 10 connections and the retention. For instance, employees with first 10 connections of low degree are about 30% more likely to leave their new companies in 1.5 years. Since the average degree of first 10 connections affects the network size later, this correlation agrees with the intuition that having a sound social network within a company might be good for increasing the employee retention.

On the other hand, the employee retention can be attributed by other factors such as seniority level. For example, employees might be able to learn from more senior employees, so having such connections can be helpful for staying longer in a company.

We study the relationship between seniority difference and turnover rate in 1.5 years. Figure 12 illustrates the negative impact of the average seniority difference between an employee and their first 10 connections on the employee's turnover rate. In other words, there is a positive correlation between the seniority difference and retention. To be concrete, employees who are connected with more than two level higher status employees are approximately 40% less likely to leave the company within 1.5 years. R^2 statistic in this analysis is very close to 1, so the correlation between the seniority difference in first 10 connections and engagement is very strong.

There are many potential scenarios to explain this observation. For instance, active employees might approach to very senior employees in the beginning and establish the relationship with them. These active employees tend to be more engaged with the companies than normal or passive employees. Another possible explanation is that as senior employees can mentor juniors for better motivations having more senior employees in first 10 connections can be helpful in the engagement with the company. At the moment, these conjectures must be left at the level of hypothesis, and we leave them to future work.

6. CONCLUSION

Understanding the growth of an individual's ego-network within a new community is valuable for common social networking problems such as link-prediction. Also, within the context of a company, the ego-network growth of a new employee can provide insights on their current position and future prospects within the company.

In this paper we investigated two common patterns that influence connection behavior in social networks, triangle-closing and homophily, using company networks in LinkedIn. the connection networks of companies on Linkedin. Based on these patterns, we demonstrated that an employee's first few connections within a new company play a significant role in forming later connections as well as even in the retention. We found that the size of an employee's ego-network after 1.5 years is positively related to the degree of their first few connections. Also, the diversity of an employee's first connections is indicative of the diversity in their total ego-network after 1.5 years. Finally, we showed that the composition of first few connections affects the turnover rate.

This paper showed many relations between the characteristics of an employee's first connections and the characteristics of their final network, and explained the relations using triangle-closing and homophily. However, as such relations do not necessarily imply any causal effect, we leave the establishment of causal relations as future work. The development of a model that can explain our observations also remains as future work.

7. REFERENCES

[1] A. Anderson, D. Huttenlocker, J. Kleinberg, J. Leskovec, and M. Tiwari. Global diffusion via cascading invitations: Structure, growth, and homophily. In *WWW*, 2015.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *KDD*, 2006.

[3] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *WSDM*, 2011.

[4] R. S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, 2009.

[5] W. C. Carter and S. L. Feld. Principles relating social regard to size and density of personal networks, with applications to stigma. *Social Networks*, 26(4):323–329, 2004.

[6] J. N. Cummings. Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science*, 50(3):352–364, 2004.

[7] P. A. Dow, L. A. Adamic, and A. Figgeri. The anatomy of large facebook cascades. In *AAAI*, 2013.

[8] R. Eisenberger, F. Stinglhamber, C. Vandenberghe, I. L. Sucharski, and L. Rhoades. Perceived supervisor support: contributions to perceived organizational support and employee retention. *Journal of Applied Psychology*, 87(3):565, 2002.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, 1999.

[10] T. H. Feeley, J. Hwang, and G. A. Barnett. Predicting employee turnover from friendship networks. *Journal of Applied Communication Research*, 36(1):56–73, 2008.

[11] S. Goel, A. Anderson, J. Hofman, and D. J. Watts. The structural virality of online diffusion. *Management Science (To Appear)*.

[12] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *EC*, 2012.

[13] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[14] P. Holme. Core-periphery organization of complex networks. *Physical Review E*, 72:046111, 2005.

[15] C.-J. Hsieh, M. Tiwari, D. Agarwal, X. L. Huang, and S. Shah. Organizational overlap on social networks and its applications. In *WWW*, 2013.

[16] X. Huang, M. Tiwari, and S. Shah. Structural diversity in social recommender systems. In *RSWeb*, 2013.

[17] M. Kim and J. Leskovec. Network completion problem: Inferring missing nodes and edges in networks. In *SDM*, 2011.

[18] M. Kim and J. Leskovec. Latent multi-group membership graph model. In *ICML*, 2012.

[19] A. M. Kleinbaum, T. E. Stuart, and M. L. Tushman. Discretion within constraint: Homophily and structure in a formal organization. *Organization Science*, 24(5):1316–1336, 2013.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *FOCS*, 2000.

[21] A. Lehman. *Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide*. SAS Press, 2005.

[22] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, 2008.

[23] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, 2007.

[24] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.

[25] J. M. McPherson, P. A. Popielarz, and S. Drobnic. Social networks and organizational dynamics. *American Sociological Review*, 57(2):pp. 153–170, 1992.

[26] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:pp. 415–444, 2001.

[27] R. Reagans and E. W. Zuckerman. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization Science*, 12(4):502–517, 2001.

[28] J. E. Sheridan. Organizational culture and employee retention. *Academy of Management Journal*, 35(5):1036–1056, 1992.

[29] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *PNAS*, 109(16):5962–5966, 2012.

[30] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[31] T. Wolf, A. Schroter, D. Damian, and T. Nguyen. Predicting build failures using social network analysis on developer communication. In *ICSE*, 2009.

[32] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike: Joint friendship and interest propagation in social networks. In *WWW*, 2011.