Network Analysis of Organizational Communication

Adam Hilgenkamp
School of Informatics and Computing
Indiana University - Bloomington

Abstract

In this paper I examine one aspect of the internal communication network of a Fortune 500 company. The company in this study uses Salesforce Chatter as a tool to communicate with other teams and with customers. Using an anonymized version of Chatter activity data I create a macro network of group membership and three micro networks to study interactions within and between groups. I start with visualizations to identify patterns in the demographics and diversity of the network. I then apply both a simple and complex contagion model to the three micro networks to show how information diffuses in the networks. The analysis is primarily done by implementing tools available in the python package networkx. The code and data to replicate this analysis can be found on GitHub.

Background

People are among the most valuable resources in a corporation. For a business to be successful it is important that its people are able to meet deadlines, report issues, solve problems and deliver quality products or services. Managers in these organizations spend considerable time and money in an effort to find the optimal organizational structure (Hanson and Krackhardt 2014). Studies have shown that the formal organizational structure is only the surface level of the organization (Duhigg 2016; Hanson and Krackhardt 2014; Tavares and Yamkovenko. 2017). The informal network that lies below this surface can provide insights that may not be apparent in the official organizational structure. One important insight we can get from studying the professional network is to understand how teams and individuals interact in a corporation. In 2016 Charles Duhigg wrote an article for the New York Times about Google's study on what makes a great team. In the study Google researchers looked at particular interactions between members of a team to understand what factors may contribute to the success of the team (Duhigg 2016; "Re:Work - Guide: Understand team effectiveness." n.d.). In addition to the connections within a team it may be helpful to understand how knowledge and advice is shared between teams (Hanson and Krackhardt 2014). A better understanding of what makes a successful team or a successful employee can help companies identify strategies to improve job satisfaction and productivity. This can potentially lead to improved product quality, reduced turnover and financial success for the corporation.

Another aspect of the organization that can be studied by observing the communication network is the diversity and inclusion in an organization. In an article written by Bogdan Yamkovenko and Stephen Tavares for the Harvard Business Review the authors looked at the decision making, idea sharing and emotional support network between men and women. They noted that there were significant differences and that simply having a diverse workforce does not guarantee true diversity (Tavares and Yamkovenko. 2017).

The Google and Harvard Business Review studies suggest that there is much to be learned from the communication network inside of an organization. In this paper my goal is to explore and observe the communication network of a specific form of communication. With these observations I will look at the communication patterns between job levels, tenure and work location. I will then look at how information may diffuse through the network. Specifically looking at how a simple contagion such as a post providing an update on a major project and a complex contagion such as recruiting team members for a volunteer event may behave in the network. With this, I aim to provide a better understanding of how Chatter is used and to identify areas for further research regarding the communication networks created with this tool.

Methods

There are many ways that employees communicate with each other and with customers. When choosing how to study the communication network, consideration was given to the accessibility of data, collection methodology and privacy of data. Considering these factors I chose to study public posts and activity on Salesforce Chatter. Salesforce is a platform with the goal of helping companies better connect with and provide exceptional experiences for their customers ("CRM 101: What is CRM?" n.d.). As a part of that platform Chatter is used to communicate with other employees, teams and customers. It is similar to Facebook or Twitter in that users can post questions, comments, react to posts and share files. Users also have the option to join groups to stay connected to information that is relevant to them ("What is Chatter?" n.d.).

To study the Chatter communication network within the organization I accessed available reports on activity, group membership and public user details. From the user detail dataset I was able to gather information about work location, tenure and job level. The group dataset contains information on group membership and the activity dataset contains comments and posts within a group. Due to the volume of the comments and posts dataset I filtered the report to only include activity in two groups over the past year.

I then downloaded all three reports from Salesforce to create a modified version of the raw dataset. First, to make the data more meaningful I created buckets for the employee attributes that I wanted to study. For work location I limited the groups to any office that had more than 50 employees. All other offices were grouped as "Other". For tenure I grouped the data by less than 1 year, 1 to 3 years, 4 to 8 years and greater than 8 years. The tenure grouping was done based on percentages of the tenure distribution (Fig. S1). Job grade was grouped by less than 4 (hourly), 4 to 6 (individual contributor), 7 to 11 (management) and greater than 11 (senior management / executive). Another important feature of this data is that it contains contractors and people who have left the company. In this case their attributes are shown as "nan". Lastly, to preserve anonymity names were replaced with a randomly generated ID. These final datasets are what was used to create the graphs used in the study of the Chatter communication network.

Using the group membership dataset I created a graph to model the macro structure of the Chatter network. In this graph I consider one particular type of interaction. Each user (u) can be connected to multiple groups (g). The way that this network is structured means that the max number of nodes would be: $n_{max} = u + g$. The result is an undirected graph connecting groups by the users that are a part of each group.

To build the interaction networks I used the Chatter activity dataset to create three separate graphs. All three graphs were created by considering the interactions between an original poster (p) and users commenting on that post (c). The original poster can also comment on their original post meaning that the interactions graphs contain self-loops. The graphs used in this analysis were unweighted and undirected graphs but future research could include studying a weighted version of the graphs. The first of the interaction graphs studies a Chatter group for one team in the organization and the interactions between users in that group. The second group is created by using activity data for a group where multiple teams go to get information about data. Lastly, the third graph is created by combining data from both groups.

The data and networks used in this analysis are good sources to observe the structure of the Chatter network but there are some potential issues that I consider when making conclusions with this data. The first challenge is in addressing the true communication structure of the organization. Chatter is not the only form of communication. Email, verbal communication, document collaboration and instant messaging also contribute to this network. When discussing the results and conclusions from this data I am only observing the communication network that manifested in Chatter. The second challenge with the data is confounding variables. Since I am just observing the network as it is the behaviors can be influenced by factors such as homophily, external and social factors. As a result users in the same Chatter groups and users that communicate frequently with each other may share similar characteristics.

Results

To better understand the general topology of Chatter in the organization I began my analysis by looking at the group membership network. The membership network consists of 71,732 nodes, 9,417 of which are groups. The remaining 62,315 nodes are individual users. Connecting users and groups there are approximately 1.53 million edges. Given this information we can calculate the average degree for the graph: $< k >= \frac{2L}{N} \rightarrow \frac{2 \cdot 1532537}{71732} = 42.73$. This suggests that the average Chatter group has approximately 43 members. However, when looking at the degree distribution it is clear that significant outliers are effecting the average degree (Fig. S2). Since the degree distribution

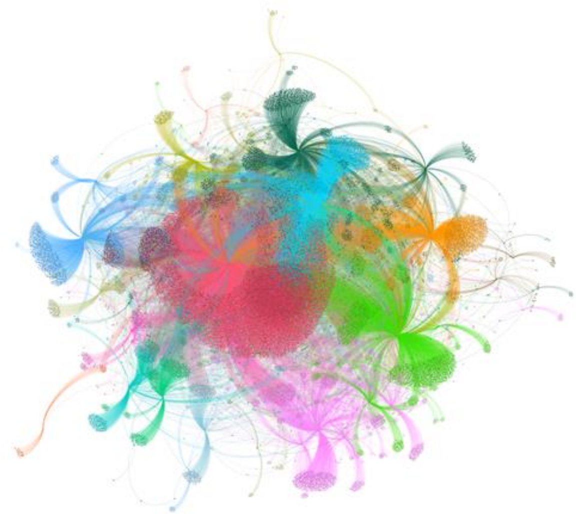

Figure 1: A visualization of a 10% sample of the group membership network. Node size represents the PageRank centrality and color represents communities detected using the modularity.

of this network appears to nearly follow the power law (Fig. S3) it is important to consider the median degree of the network as well in order to get a better understanding of the typical group or user. The median degree of the Chatter group network is 14 and the upper quartile is 40. Both mean and median together provide a more holistic understanding of the macro scale of the Chatter communication network.

In addition to observing the network statistics I wanted to visualize the Chatter group network in order to get a sense of the topology. Given the size of the network I choose to take a random sample of 10% of the chatter groups and connect the members of those groups (Fig. 1). To help identify the hubs in the network I determined the node size based on the PageRank centrality measure. I also used the modularity with a resolution of 0.8 to identify communities in the network. When studying the resulting visualization I found that the large groups tend to be the most central with many small groups loosely connected to the central part of the network. This was not a surprising finding as we would expect the largest groups to be the most central. The interesting observations were when I looked closely at the nodes connecting groups. One example is a group for the Japan sales team and the connections to a group whose main focus is to discuss and share information about a particular product. There are connections between the groups but there is a significant portion of the sales group that is not a member of this particular product group. Further study of this fact may allow quantification of the importance of being connected to information about a product. If it is possible to find a significant difference in the sales of that product between the users connected to the product Chatter group and those who are not it may provide useful data for sales managers.

Group membership expresses the macro scale of the Chatter network but there is another layer within and between these groups. To study this layer I looked at the three different interaction networks. The first interaction network is composed of posts and comments between members of a group for one particular team in the organization (I will refer to this as the team group). The interaction network for this group has 89 nodes and 157 edges. The average degree for the graph is: $< k >= \frac{2L}{N} \rightarrow \frac{2 \cdot 157}{89} = 3.53$. The weighted degree for the network is:

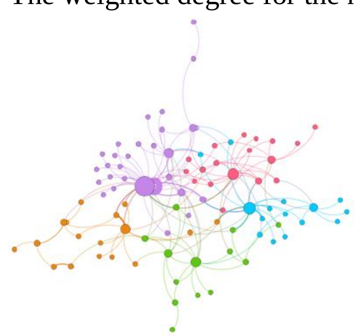

Figure 2: A visualization of a single Chatter group and the interactions within that group. The node size is set using the PageRank centrality and the color represents the communities identified by the modularity.

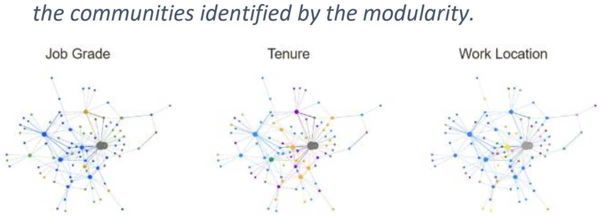

Figure 3: Visualization of attributes in the network for the team group.

$< kw >= \frac{1}{N}\sum_{i=1}^{n} kw_i \rightarrow \frac{1}{89} \cdot 436 = 4.9$. The weights are derived by counting the number of posts or comments between two users. When plotting the degree distribution of the network it also appears to generally follow the power law (Fig. S4). This means that a small number of group members are contributing heavily to the activity. To describe how many interactions the typical user has the median degree of 2 may be more accurate given the outliers in the distribution.

In the interaction networks I also studied the clustering and distance measures. For the team group the average shortest path was 3.09 and the diameter was 7. This means that on average 3 steps should connect one user with another but in some cases the longest shortest path could be 7. The average clustering coefficient of this network is 0.144 which is similar to the clustering coefficient found in Facebook and Twitter networks (Ch'Ng, 2015; Ugander and Karrer, 2011).

To look for patterns in the network for the team I created a visualization using Gephi (Fig. 2). To visualize centrality I used the PageRank centrality measure. This is represented by the node size in the visualization. To identify communities in the network I used the modularity with a resolution of 1.5 which detected 5 communities in the network. I then took the visual analysis a step further and looked at the different user attributes that I added from the user detail dataset. First, I looked at how work location factored into the network interactions. When studying the interactions it appears that the locations are well mixed in the network even though it appears that this team has a large presence in one location (Fig. 3). I then looked at tenure to see if there were any patterns in users interacting with others who have been with the company a similar amount of time. This was again well mixed which suggests that tenure is not a major factor in communication. Lastly, I looked at job grade which is grouped by hourly employees, individual contributors, mangers, and executives. Job grade also shows a higher number of managers in the group but the results are roughly close to the population distribution for job grade (Fig. S7). The interactions between job grade don't appear to be isolated suggesting that like location and tenure, job grade may not be a major factor in communication patterns in the group. A deeper statistical analysis would be necessary to support these assumptions.

The second interaction network shows the connections between users in a group where employees go to ask questions about data (I will refer to this group as the data Q&A group). This network consists of 283 nodes and 883 edges representing the posts and comments between users. The average degree for the graph is: $< k >= \frac{2L}{N} \rightarrow \frac{2\cdot 883}{283} = 6.24$. The weighted degree for the network is: $< kw >= \frac{1}{N}\sum_{i=1}^{n} kw_i \rightarrow \frac{1}{283}\cdot 4580 = 16.18$. As was done with the single team group the weights are derived by counting the number of posts or comments between two users. This group also follows the power law but does display a tail which deviates from the theoretical power law (Fig. S5). The median of the degree distribution is 2 which is the same as the median for the team Chatter group. Since the weighted and unweighted average degree is higher but the median is the same it suggests that the outliers in the data Q&A group are more significant. The average shortest path for this group is 3.04 which is almost the same as team group. The longest shortest path in the network is 6 which is slightly smaller than the diameter of the team group. The average clustering coefficient for the data Q&A group is 0.21 which is higher than the team group.

To study patterns in the network I applied the same visualization methods that I used for the team group to the data Q&A group. The node size is determined by the PageRank centrality and the community structure is set by the modularity with a resolution of 1.5 (Fig. 4). I then looked at the work location, tenure and job grade interactions in the network. A significant portion of this group is located at one work location (53%). This varies from the total population, where 20% of users are located at the same work location (Fig. S6). Even though many of the users of this group are located in one location the central nodes in the group are not all in one location. From observation work location seems to be diverse across the data Q&A group (Fig. 5). When looking at tenure I observed similar behavior. The distribution of the group roughly follows that of the population and the important users span across tenure groups. Lastly, when studying job grade in the data Q&A group I found that the individual contributors and the managers are well mixed but there was only one executive that had interacted in the group. Based on this observation it appears that executives are relying on their teams to deliver information about the data that they are using to make decisions. This finding was not entirely unexpected but looking at group interaction in this way may highlight the areas that are top of mind for executives.

Figure 4: A visualization of the data Q&A group. The node size is set using the PageRank centrality and the color represents the communities identified by the modularity.

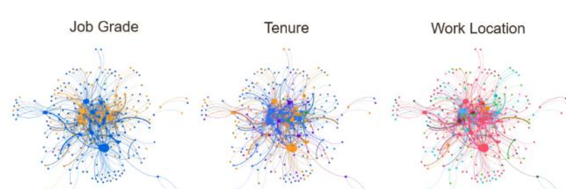

Figure 5: Visualization of attributes in the network for the team group.

After developing an understanding of each individual group I studied how the dynamic might change when combining the groups together and how information diffusion may be different when combining each group. The combined group consists of 368 nodes and 1040 edges. The average degree for the combined graph is 5.65 and the average weighted degree is 13.63. The decrease in the average weighted and unweighted degree is expected as the team group with a smaller average degree will lower the combined group average. The combined group also follows the power law distribution and looks similar to the distribution for the data Q&A group (Fig. S8). The average clustering coefficient and the density of the combined graph also reflect the measures for the data Q&A group. The interesting difference is that the diameter increases to 8 and the average shortest path increases to 3.73 (Table 1). This is likely driven by the introduction of weak ties between the Chatter groups.

To visualize the connections between groups I followed the same process implemented for the team and data Q&A group. I used PageRank centrality for the node size and modularity with a resolution of 1.5 to identify communities (Fig. 6). The orange community represents the team group and the larger component is the data Q&A group. When looking closely at the combined network we can see that there are only two connections between the groups. Given the small number of connections between groups there could be effects on the efficiency and success of the team. If one or both of the key users leaves the company the connection to the data Q&A group could be broken. As a result other forms of communication will be necessary and important information regarding data may be missed.

To take the analysis further I then applied two different information diffusion methodologies to test the spreading capabilities of the combined network. First, I looked at simple contagion. The SI epidemic model is a good approximation for simple information spread (Barabási, n.d.).

Figure 6: A visualization of the combined groups. The node size is set using the PageRank centrality and the color represents the communities identified by the modularity.

Statistic	Team Group	Data Q&A Group	Combined
Nodes	89	283	368
Edges	157	883	1040
Avg Degree	3.53	6.24	5.65
Avg Weighted Degree	4.90	16.18	13.63
Median Degree	2	2	2
Friendship Paradox	0.89	0.92	0.91
Density	0.04	0.02	0.02
Diameter	7	6	8
Avg Shortest Path	3.09	3.04	3.73
Avg Clustering Coefficient	0.14	0.21	0.20

Table 1: Table comparing the statistics for each group individually and combined.

An example of simple information could be a post regarding an upcoming update to a dimension in a commonly used dataset. One person will make the original post and given the probability that a user will see the post we can make some assumptions of how that information will spread in the network. To test the combined Chatter network I used an implementation of this model in python (Ahn, YY & Elise Jing, 2018). The beta in the model represents the probability that the infection or information will spread. To test the characteristics of the combined group I chose a beta of 0.9 giving a 90% chance that a neighbor in the network would see the information. This choice is just an approximation and further study would need to be done to see how likely it is that a user sees a post. I also chose to set the initial infection to 1. This represents the single user making the post. I then ran the model for each Chatter group separately and then ran the model for the combined group. To compare, I also ran the model on an Erdos Renyi random graph with the same number of nodes and connection probability and on a configuration model using

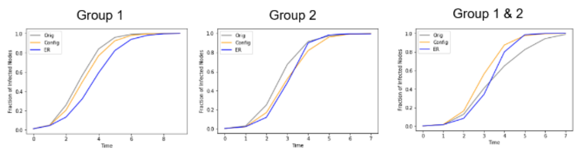

Figure 7: Graph of "infected" nodes over time. Generated using an SI model with a beta = 0.9 and an initial infection of 1 node.

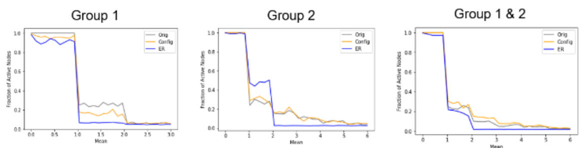

Figure 8: Graph of "infected" nodes given different thresholds. Generated using a threshold model.

the real graph's degree sequence. I found that the real network performed slightly better than both random models when run on the individual Chatter groups. However, when running the model on the combined chatter group the infection or information spread at a lower rate (Fig. 7). In this situation the introduction of weak ties when combining the groups acts like a bottleneck for the spread of information. Since there are only a few connections the likelihood that the contagion spreads across those edges is lower.

In addition to the simple contagion model I studied what would happen with a more complex contagion. An example of this type of information would be creating a post asking co-workers to sign up for a volunteer event. For a particular user to take this action they may require two or three of their neighbors to sign up before doing so themselves. A good model for testing this behavior is the threshold model (Granovetter, 1978). I applied a python implementation of the threshold model to test the behavior of the Chatter groups and to compare the results against comparable random graphs (Ahn, YY & Elise Jing, 2018). The results of the test on the team group showed that if the mean threshold is between one and two the amount of infected nodes will be approximately 25%. Once the threshold passes two there are effectively zero infected nodes (Fig. 8). The real graph did perform better than both random graphs but the low clustering coefficient contributes to the behavior given a threshold. The data Q&A group has a slightly higher clustering coefficient and it showed spreading behavior past a threshold of two. The configuration model and the real model exhibited similar behavior and showed that spreading could take place even to a mean threshold of 5. The Erdos Renyi model had a higher infection rate between 1 and 2 but went to zero past a threshold of two (Fig. 8). When the data Q&A group was combined with the team group the behavior of the complex contagion did not seem to change significantly. The updated parameters effected the Erdos Renyi model but the behavior of the real graph and the configuration model was comparable to the data Q&A group on its own.

Conclusion

In summary, I looked at how networks form in Chatter. Similar to how tools like Facebook or Twitter are used in the social world; Chatter facilitates interactions between users in a business context. Understanding how tools such as Chatter are used can help illuminate the real structure of the business beyond the formal organization chart. My analysis of Chatter is an initial exploration of this data but further analysis could highlight important facets of organizational communication and inclusion. In my analysis I found that on a macro scale there are many Chatter groups that have a small number of members but there are a handful of groups that almost every user is a part of. Future research to understand how group membership may contribute to sales performance or other performance measures could provide value to companies and teams. I then studied the characteristics of two individual groups and looked for patterns in the attributes of the users in those groups. The groups that I chose to study did not show any major isolation of a particular group. I then looked at how the dynamics of the network change when you combine the groups and introduce the weak ties that hold them together. With this lens I looked at how information may spread in each group individually and how it might change as groups are connected together. In future research I would like to look at more data and more groups to expand the scope of the study. The data that I collected focused on posts and comments but to get a more holistic picture I would like to include data on views and likes. I also think it is important to combine this data with an analysis of other types of communication such as email, survey data and instant messaging. There are certainly privacy concerns that go along with this type of analysis. I made sure that the data used did not include any of the actual posts and also removed any personally identifiable information from the data. This may limit to what extent this type of analysis can be applied in a business setting. At a high level observations and changes can be made to make the network more effective such as encouraging employees generally to join particular groups. However, it may be ethically challenging to specifically ask someone to join a group because their sales numbers for a particular product are behind. In addition, there are challenges with understanding the impact of external pressures and homophily. Some managers may require that their team joins a group while others may not. Users will also self-select into groups with people that are similar to them. Still, Chatter data and the analysis of it can still produce interesting and useful information. This information can provide a window into the informal structure of the business. The interactions on Chatter can also provide another way to measure diversity that is deeper than the percentages of the employee population. Effective communication in an organization is key to its success. Using the tools of network science we can better understand the current structure and find ways to improve that structure to achieve business growth, increase efficiency and to empower employees.

References

1) Barabási, A. (n.d.). Chapter 10: Spreading Phenomena. Retrieved May 2, 2019, from http://networksciencebook.com/chapter/10
2) Ch'Ng, E. (2015). Local Interactions and the Emergence of a Twitter Small-World Network. Social Networking, 04(02), 33-40. doi:10.4236/sn.2015.42004
3) C. Duhigg, (2016, February 25**)**. What Google Learned From Its Quest to Build the Perfect Team. Retrieved February 12, 2019, from https://www.nytimes.com/2016/02/28/magazine/what-google-learned-from-its-quest-to-build-the-perfect-team.html
4) Granovetter, M. (1978). Threshold Models of Collective Behavior. American Journal of Sociology, 83(6), 1420-1443. doi:10.1086/226707
5) J. Hanson, D. Krackhardt. (2014, August 01). Informal Networks: The Company Behind the Chart. Retrieved February 12, 2019, from https://hbr.org/1993/07/informal-networks-the-company-behind-the-chart
6) S. Tavares, B. Yamkovenko. (2017, September 20). To Understand Whether Your Company Is Inclusive, Map How Your Employees Interact. Retrieved February 12, 2019, from https://hbr.org/2017/07/to-understand-whether-your-company-is-inclusive-map-how-your-employees-interact
7) CRM 101: What is CRM? (n.d.). Retrieved from

https://www.salesforce.com/crm/what-is-crm/
8) What is Chatter? (n.d.). Retrieved from https://www.salesforce.com/products/chatter/overview/
9) Re:Work - Guide: Understand team effectiveness. (n.d.). Retrieved February 12, 2019, from https://rework.withgoogle.com/guides/understanding-team-effectiveness/steps/introduction/?gclid=Cj0KCQiA-qDTBRD-ARIsAJ_10yInGFDkECEnHtYYehM8VW1Ab_uAdMd0gzOhQh58RFJu2qURtWTeZOYaAi3REALw_wcB
10) Ugander, J., & Karrer, B. (2011). The Anatomy of the Facebook Social Graph. Retrieved May 2, 2019, from https://arxiv.org/pdf/1111.4503.pdf.
11) Ahn, YY & Elise Jing (2018). Epidemics on networks: the SI model. https://github.com/yy/netsci-course/blob/master/m10-epidemics/epidemics_assignment.ipynb
12) Ahn, YY & Elise Jing (2018). The Threshold model. https://github.com/yy/netsci-course/blob/master/m12-diffusion/threshold_model_assignment.ipynb

Appendix

GitHub Repository:
https://github.com/ahilgenkamp/Network_Sci_Project

Supplementary Figures & Tables

```
25% --> 1.0
50% --> 2.0
75% --> 4.0
95% --> 8.0
```

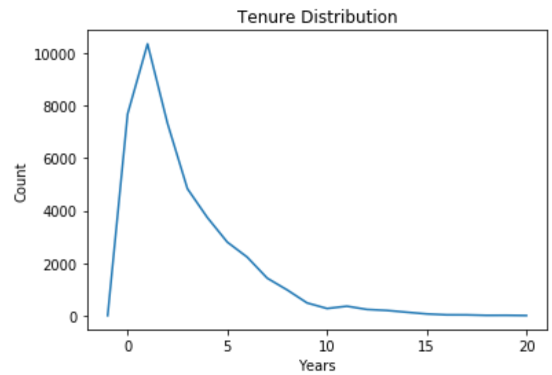

Figure S1: Distribution of tenure in the user dataset. Binning of tenure was based on percentages of the tenure distribution to show more meaningful groups when visualizing tenure.

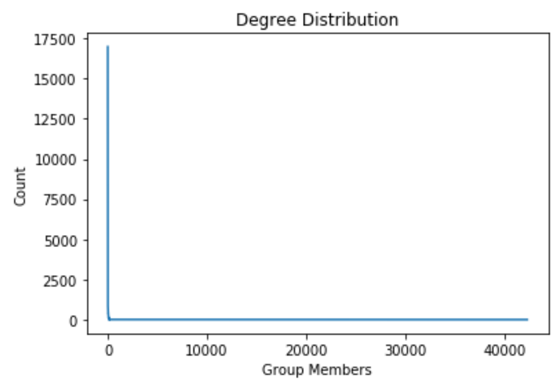

Figure S2: The plotted degree distribution of the Chatter group membership network.

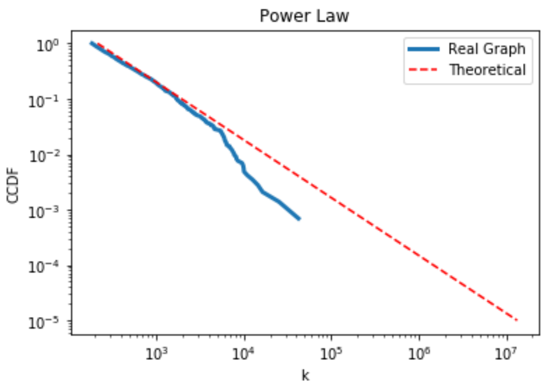

Figure S3: The plotted degree distribution of the Chatter group network vs a theoretical power law distribution.

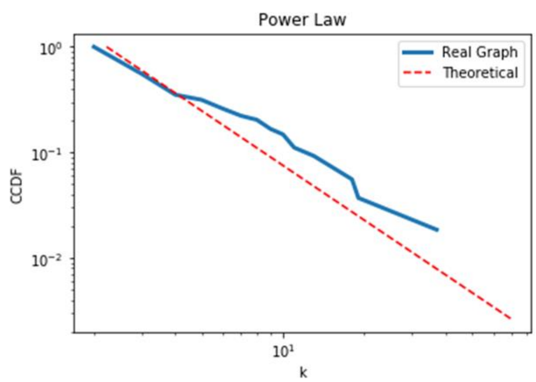

Figure S4: The plotted degree distribution of the single team Chatter group interaction network vs a theoretical power law distribution.

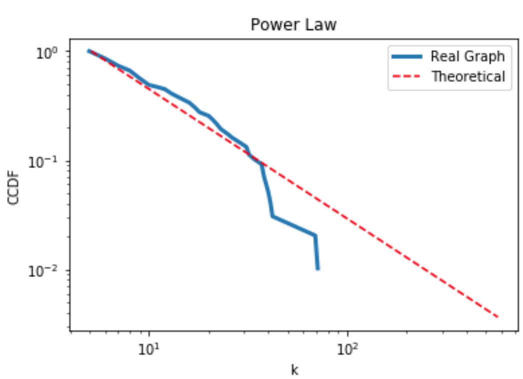

Figure S5: The plotted degree distribution of the data Q&A Chatter group interaction network vs a theoretical power law distribution.

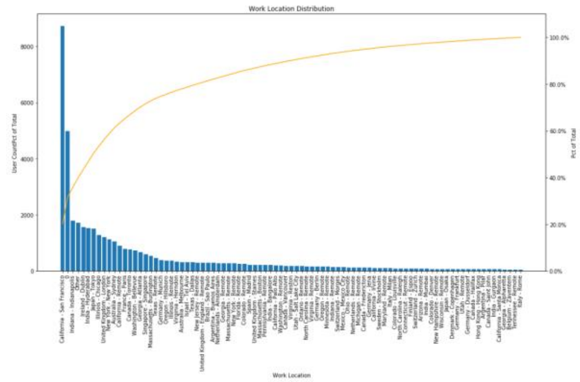

Figure S6: Combo chart showing count and the percent of the total for each office location for the entire population.

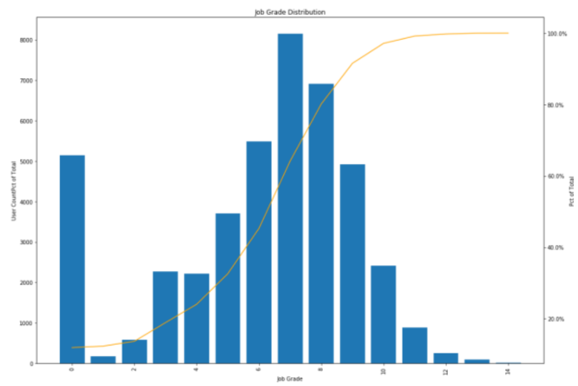

Figure S7: Combo chart showing the job grade distribution of the total user population.

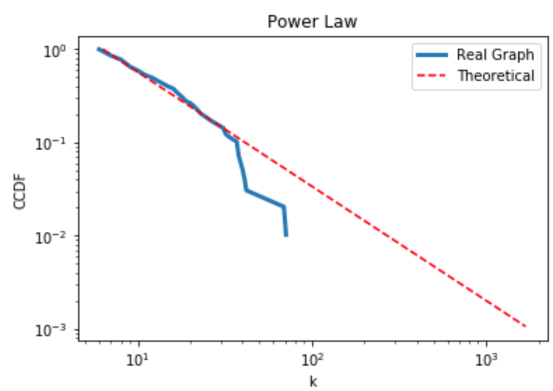

Figure S8: The plotted degree distribution of the data Q&A Chatter group interaction network vs a theoretical power law distribution.