Which null models most accurately model social circles in real life ego networks?

David Mora and Rachel Gianforte
dmora1, rachelg4 @stanford.edu

Introduction

Research surrounding the explanatory and predictive power of social networks has exploded in the last decades: wide ranging research reveals a person's social network links inextricably to their susceptibility to mental illness[1], obesity[2], or even of committing violent crimes[3]. In these widespread applications, ranging from policing to epidemiology to health, existing research has only begun to elucidate the uses of social network analysis. Consider for example, medicine, where, currently, disjoint checkboxes and static files merely outline *family* health history, where medical practitioners would be better served by examining a network of social and familial health, placing individual symptoms within a holistic context and drawing comparisons across patients[2].

However, as it stands, network analysis of individual's social network remains largely in the niche sector of academic research and multi-million, longitudinal studies, and widespread clarity on paradigms of analysing individual social circles largely does not exist.

The opportunity space, however, is changing.

Coincident with the rise of network research and applications, the world's largest and unprecedented documentation of social networks, Facebook, grew to over 1.7 billion active users, among many other expansive networks like Google+ or LinkedIn. These largely pubic sources are saturated with information about users and their relationships to others. How might the these social networks be harnessed toward network-based analysis?

Examining the current literature, while summary statistics of overarching network structures [8], general detection of circles within ego-networks[6], and examples of the predictive power of social network analysis[9] exist, algorithms or paradigms with which to specifically irrigate the saturated social information of ego-networks are still emerging, more mature forms of which are needed to form the basis for understanding and comparing individuals in a network.

In our project, we aim to build off previous research on community detection and graph similarity scoring toward laying the groundwork for characterizing social-media documented ego-networks with labeled social circles. Specifically, we first attempt to find structural trends based on feature-specific circles both within ego-networks. Upon finding the structure of social circles to be largely random and not correlated with node features, we then explore this random nature by testing the Facebook data against various null models, ultimately designing and testing our own. Because of the typically small nature of social circles, we provide visualizations of networks to aid in intuition.

While this paper forms just a part, the larger applications of understanding social circles in ego networks, we think, are powerful and far-reaching: social network analysis can now move from overall network properties, average node descriptors or niche discussions toward ego-network optimized analytical methods that sheds direct light on individual social circles. For example, given just a few ego networks in a larger graph, researchers would be able to use our ego-network building model to fill in a synthetic, larger network, and then model other phenomena such as information or virus spread. As another example, medical professionals could use this model to build and analyze the ego networks of patients despite incomplete datasets.

Understanding the nature of real-world social networks on a more granular level holds potential for a vast number of fields centered around human interaction, organization, and wellbeing.

Literature and Prior Work

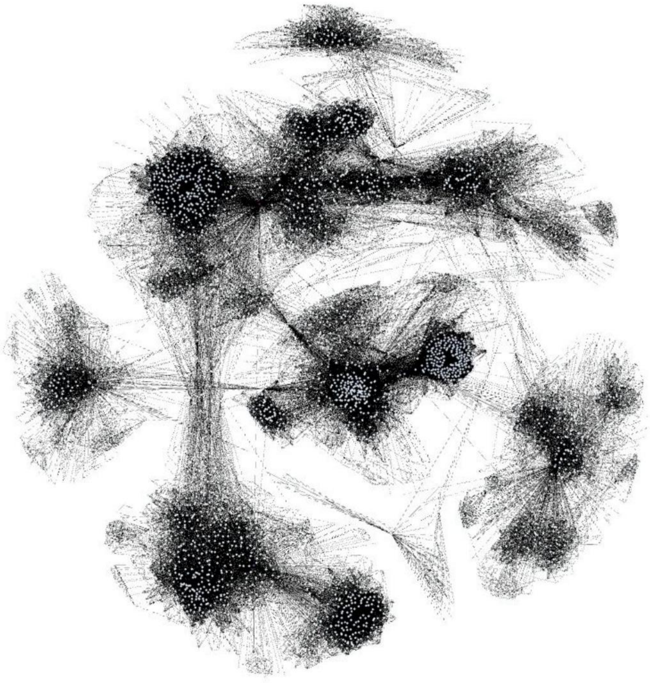
Substantial research exists surrounding dividing large network into clusters of social circles as well as studying the structure of networks overall. For example, the paper "The Anatomy of the Facebook Social Graph" provides an in depth study of the structure of the overall network[4]. It looked at trends in the overall network and the discovered that strong communities exist. Sampling from nodes to perform global computation, the authors also found global modularity largely based on geography, acute age-based assortative mixing, highly clustered cores of friends in an overall very sparse graph, and a stunning "small world" where the average distance between any two users was 4.7 hops. Their work began much of the research that now focuses on the very large graphs we now have access to through mass social media networks, but offers no ego-node resolution insight.

Leskovec and Handcock et al have both focused on how to organize social networks into clusters of friends[6][7]. Hancock tested both a maximum likelihood estimation method and a Bayesian method to determine how many clusters there were and which nodes belonged to which clusters. Leskovec, however, sorted nodes into clusters based on similarity in features. Both papers resulted in similar overlap between circles/networks even though circle generation within a network differed greatly. These insights harken toward an underlying structure in how social networks form and overlap [7].

Arnaboldi et al examine frequency of ego-alter interaction on Facebook, and reveal that online social networks display high similarity to real-world social networks: both seem to follow Dunbar's circles, rings of intimacy which grow in size as the decrease in intimacy [10]. This paper suggests that studying online social networks connects closely to existing research on social networks, but leaves a gap in discussing and characterizing circles within ego networks, in large part because the study lacked access to a ego graphs with hand labeled social circles.

Newman and Girvan used iterative edge removal to measure robustness features of social networks, but did not study the clusters' characteristics once detected [5]. They also developed an algorithm to test the strength of a community using the number of edges between nodes in the community divided by the number of from the community to the rest of the graph, which informs our starting investigation into characterizing communities in ego networks.

Dataset



Avg Degree	43.7
Avg Path Length	4.3
Avg Clustering Coefficient	0.6

Figure 1: Visualization of entire Facebook graph and table of summary properties

We've focused investigation on the graphs of 10 Facebook ego networks -- 4,039 nodes and 88,234 edges in a giant weakly connected component -- compiled in Jure Leskovec's research on clustering.

This overall graph dataset consists of 10 hand-labeled *ego networks* -- a network consisting of all the connections between an *ego*'s immediate neighbors, *alters,* with the ego node removed. Ego nodes are removed because they would often obscure the analysis of their social circle, ie with ego nodes still in a network, the maximum possible distance between any nodes would be two hops. Evident from the overall graph visualization, these ego networks overlap; however, our analysis focuses on examining individual ego networks.

Each node in a network is hand-labeled with features, such as what school a person went to, with the name of the school anonymized.

Before looking at specific ego networks, we compared its average clustering coefficient of 0.606 against a null model generated for it via edge rewiring, averaged over multiple iterations, which produced an average clustering coefficient of 0.054, over an order of magnitude smaller. Needless to say, like many real-world networks, the dataset displays ultra high levels clustering. We keep this figures in mind as we zoom into ego networks and circles within them.

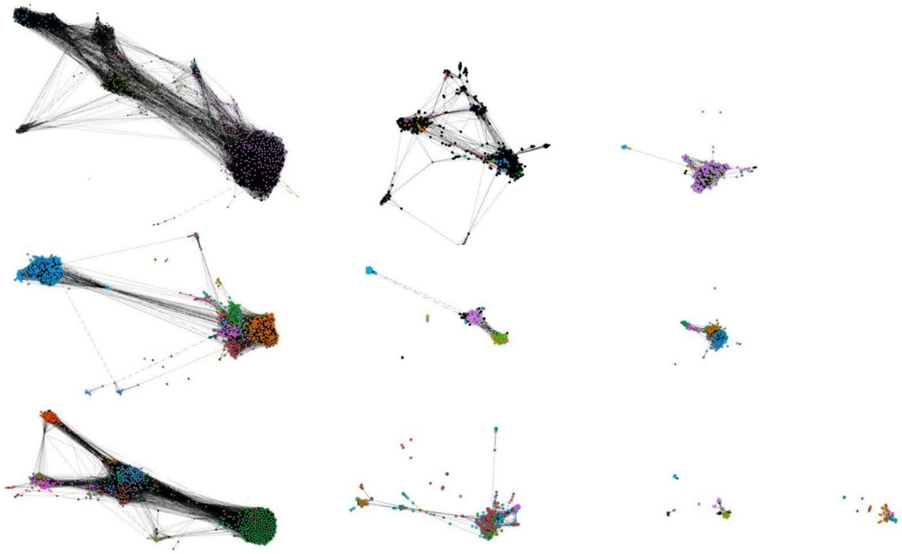
Initial Summary Analysis of Ego Networks

To begin our overall analysis and determine what type of aspects of ego-network would be interesting to study in more detail, we ran summary statistics for each of the 10 ego networks in the Facebook dataset, and then determined the mean, standard deviation, and range across them all.

	Standard Deviation	Mean	Range
Number of Nodes	1068	732	3911
Number of Edges	25049	15749	88010
Clustering Coefficient	0.081	0.568	0.271
Mean Degree	10.5	15.0	37.4
Max Centrality	0.106	0.330	0.310
Max Closeness	0.058	0.496	0.187

Chart 2

These number demonstrate that individual ego networks vary wildly. That the ego networks vary in size and structure, though certain elements such as clustering coefficient and closeness remained relatively more constant. A visualization of all 10 ego networks confirmed this (see Visualization 1, and Appendix for detailed explanation of visualization techniques).

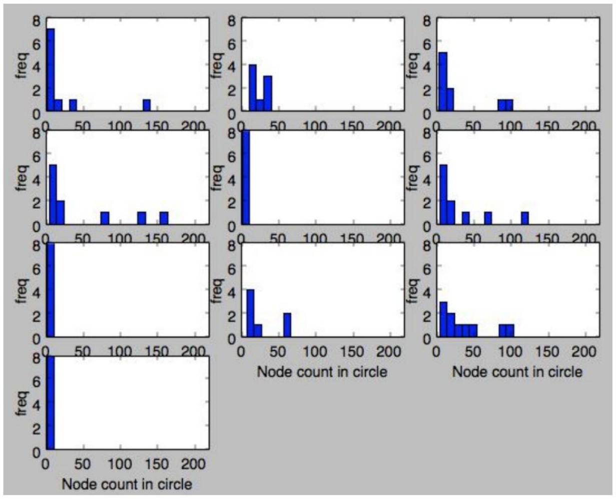


Visualization 1: Ten Ego Networks, nodes colored according to hand-labeled circles

Further, analysing the same properties across all the circles in all the ego networks, we found similarly huge variation in the basic statistics of circles (see Chart 1 in appendix)

Based on the wide variation in the egos networks, we hypothesized that circles were structured based on features rather than the ego node involved. To analyse circles by feature, if more than 80% of the nodes of a circle possessed that feature, we analyzed that circle along with others possessing that feature. However, we found no useful results. Many circles were not even associated with any feature, but displayed a mix of features. Even the circles that possessed the same feature varied as widely as circles associated by ego network.

Based on this we concluded that the size and structure of the circles we not dependent on the people involved or the features that defined them. We did hypothesize that there were many small circles (which is why the means were small), but that there were a few very large circles that skewed the statistics for all our analysis (which is why the standard deviations were large). A histogram of size of each circle, individually for each ego and cumulatively, answered this hypothesis (see Figure 1A and B).



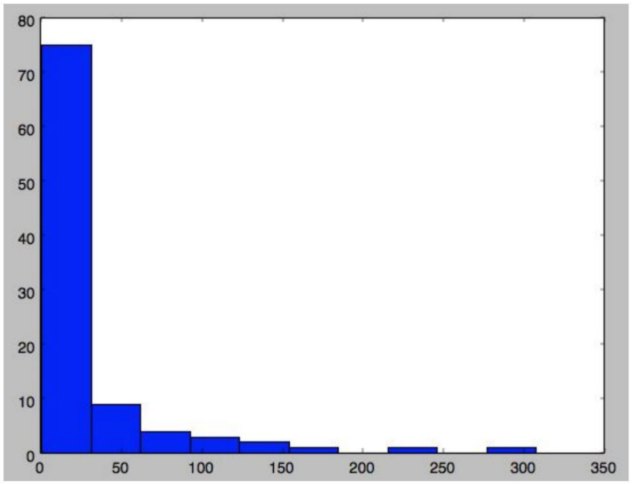


Figure 1A (left): Node-count-per-circle histogram plotted separately for each ego (note: values above 220 nodes were cut off for readability)
Figure 1B (right): A linear histogram of the nodes-per-circle, cumulative for all ten ego networks.

By studying the data we determined the cumulative histogram of circle sizes follows a power. We tested this theory by fitting a regression line to the data following the power law. We took the log of x and the log of y to allow us to fit a linear line rather than an exponential equation. We ended up with the equation

$$log(x\ number\ of\ nodes) = 9.165 - 1.79 * log(probability\ of\ circles\ containing\ more\ than\ x\ nodes)$$

This equation fit the data extremely well with total squared loss of 3.2186 over all the data points and we went on to use it to generate our null model.

Interpreting Initial Results

While this histogram confirmed our third hypothesis and uncovered a power law distribution, it also presented the issues that circles appeared much more random than we originally thought they would be. This forced us to reexamine out research and approach the problem differently.

Generating null models of ego networks with social circles

We were very interested in the varying sizes and structure of circles within ego networks, and potentially random characteristics of them, and so wanted to research if you could generate a similar structure using a null model. Past research has focused on sorting friends into circles, but we wanted to see if we could develop a random model that generated an ego network with a circle structure similar to the real world Facebook data.

In order to compare the null models to the real world facebook data, we applied several different null models to each of the facebook ego networks, generating a model for each of the ten ego networks. Because the ego networks varied so much in number of nodes and number of edges, we believed it was important to capture that variation as we evaluated possible null models.

We then recorded the standard deviation and mean for each statistic by the type of null model. We looked at statistics including number of nodes and edges per circle, closeness, and community strength. Ultimately, we focused on testing Erdos-Renyi and Small World generation, though we also evaluated rewiring edges randomly while maintaining degree distribution and preferential attachment model. Both of the last two provided very similar results to the Erdos-Renyi model.

For the Small World model, we ran the graph generation with varying values for the rewiring probability, the optimization pointing to 0.4 as best fit. This number makes intuitive sense: under it clusters could form but nodes were not so densely connected that they connected across clusters, mixing and eliminating distinct circles.

Once we generated random graphs, we had to separate them into circles. Original plans to use Professor Leskovec's method for separating an ego network into circles proved impossible since his method relied on attributes to inform his clustering, and our null models do not generate random attributes.

Instead we used the algorithm developed by Newman and Girvan -- a reductive method which iterated over the graph and calculated the betweenness values for every edge, and then removed the edge with the highest betweenness, repeated this process until the graph was separated into strongly connected components, each of which represents cluster.

This algorithm was relatively straightforward to write, but we struggled to determine when the clusters had been sorted out. Originally we tried simply running the algorithm for a particular number of iterations, but because of the variation in the sizes of the graphs this was impossible. Some graphs ended up split into individual nodes, while large graphs remained as one cluster. We also tried determining when to stop based on a metric for community strength.

However, this metric was calculated by the number of edges with thing the cluster divided by the number of total from that cluster to the rest of the graph. Though this metric is useful for analyzing clusters once they are formed, it could not be used to form measures because the metric for community strength proves largest always when there are the fewest components.

Finally we determined from our initial findings that the ego networks in the facebook data generally had ten circles, and so we stop when the number of strongly connected components equaled ten since we were generating null models based on the facebook ego networks.

Results of Null Models

	Number of circles per ego network	Number of nodes per circle	Number of edges per circle	Clustering coefficient	Betweenness Coefficient	Community Strength
Facebook	9.57	22.72	206.87	0.343	1137.22	0.629
Erdos-Renyi	6.86	31.71	575.77	0.032	192.66	0.140
Small World (.4)	9.29	23.41	146.78	0.439	259.01	0.402

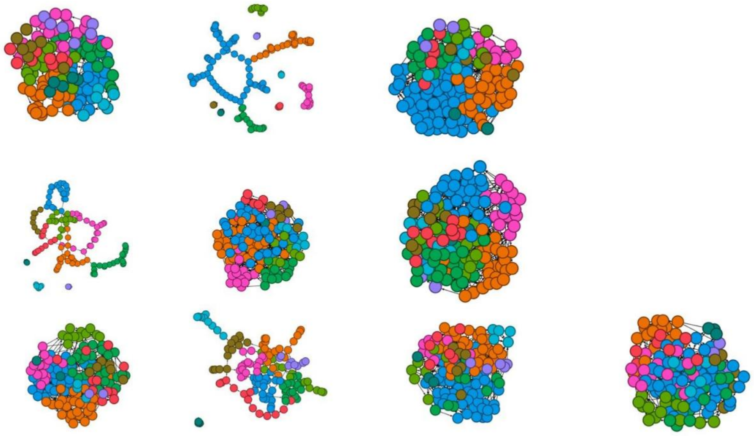
It is clear that that the null models fail to capture the circle structure of the real data set. These results, however, make sense. Since the Erdos-Renyi graph is a random generation model, it does not account for the increased probability of friends having mutual friends in the real world. Edges have probabilities of existence, but they are not based the existence of other nodes. The small world model did much better, though it too fell short. This is because the small world increases the probability that two nodes will share an edge if they are both connected to another node. This increases the clustering coefficient, but more importantly for use, also created more of a circle structure within the small world model network.

Designing Our Own Null Model

Though the small world network performed better than the other null models, we thought we could improve upon it. We realized that the ego networks were essentially made up of smaller networks (circles) , so we decided to develop our own null model that better represents the facebook ego networks.

To begin, we calculated the power law for the size of a circle using our observed data from the ego networks sorted by Professor Leskovec. We continued to use that premise that each ego network would have 10 circles. For each of the ten circle we randomly generated a number of nodes in that circle based on the power law. We used that number of nodes to generate a small world graph for that circle. We discussed experimenting with different small world rewiring probability values, but given that we were not able to study the structure of the nodes themselves, this seemed pointless. Instead we focused on generating a similar community strength. Once we generated each of the circles, we combined them into a single larger graph, and then ran the rewiring algorithm. However, we only ran it for as many iterations as 20% of the the number of edges. Since the each iteration re-wires two edges, this would lead to 40% of edge being rewired, which would generally return the community strength we were hoping to replicate.

	Number of circles	Number of nodes per circle	Number of edges per circle	Clustering coefficient	Betweenness Coefficient	Community Strength
New Null Model	10	12.78	24.11	0.233	139.24	0.571



Visualization 2: Ten Generated Ego Networks from our null model based off each of the ten ego networks from the Facebook data set

Conclusion

There is a lot more research to be done on the structure of circles. Since past research has established multiple algorithms for sorting an ego network into circles, we were able to focus on the structure of the circles. Though we were not able to determine the effects or attributes on the formation or structure of a circle, we did discover interesting things about circles in general. It would be interesting to use more data to confirm our hypothesis about the size of the circles being governed by the power law. However, we were able to develop an original null model that mimicked the structure of the circles. The key point was recognizing that each circle acts as a mini graph within the ego network, so to construct and ego network it is best to construct mini networks, and then randomly rewire edges between them to form the ego network as a whole. It would be interesting to compare our model to other social networks with circles, such as Google+.

References

[1] Kawachi, Ichiro, and Lisa F. Berkman. "Social ties and mental health." Journal of Urban health 78.3 (2001): 458-467.
[2] Christakis, Nicholas A., and James H. Fowler. "The spread of obesity in a large social network over 32 years." New England journal of medicine 357.4 (2007): 370-379.
[3] Shaabani, Elham, et al. "Early Identification of Violent Criminal Gang Members." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.
[4] Johan Ugander,Brian Karrer, Lars Backstrom, Cameron Marlow. The Anatomy of the Facebook Social Graph. Arxiv 2012.
[5] Mark E J Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2), 2003.
[6]J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.
[7] Mark Handcock and Adrian Raftery. Model-Based Clustering for Social Networks. J.R. Statist Soc A:301-354, 2007.
[8] Laura A Zager and George C Verghese. Graph similarity scoring and matching. Applied mathematics letters, 21(1):86–94, 2008.
[8] Facebook Graph Anatomy
http://snap.stanford.edu/class/cs224w-readings/backstrom12fb.pdf
[9] J. Leskovec, D. Huttenlocher, J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. In Proc. WWW, 2010.
[10] Arnaboldi, Valerio, et al. "Analysis of ego network structure in online social networks." Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international conference on social computing (SocialCom). IEEE, 2012.
[Gephi Visualization Software] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Appendix

Ego Network	Circles count	Avg num nodes	Avg num edges	Avg Clust. Co	Avg Degree
1	24	38.7	279.3	.313	2.21
2	9	89.8	2672.4	.189	8.33
3	17	67.2	1555.5	.330	6.47
4	32	56.41	1575.4	.338	11.18
5	14	2.38	1.5	0	0.237
6	17	34.39	666.3	.250	5.24
7	46	1.85	3.07	.405	0.627
8	13	21.13	296.9	.174	4.606
9	14	32.05	354.4	.281	3.610
10	7	3.29	13.2	.378	1.234

Chart 1: Circles by Ego Network

Chart 1: Circles by Ego Network

Visualizations were created using the open source application Gephi using the Force Atlas layout algorithm, which was designed by Mathieu Jacomy specially to display "Small World" and Scale-Free networks for visual analysis, while minimizing layout bias.