Memetic Tribes in Twitter Mutuals Network

Francisco Carvalho, João Andrade, Paulo Dias
Professor Francisco Santos
Instituto Superior Técnico
01/12/2019

Abstract

In this paper we follow up on the exploration of our *Memetic Tribes dataset*, undertaking three tasks. 1. We demonstrate the fidelity of a method we call **pluri-ego-sampling**, the one we used to obtain the dataset presented in our first project. 2. We try to explain the **broken power law** degree distribution exhibited by our dataset. 3. We attend to the original objective of the project, using **community-finding** methods to empirically detect the memetic tribes predicted and in *Limberg, Barnes* [6], with encouraging results.

Introduction

In this paper we follow up on the exploration of our *Memetic Tribes dataset*, undertaking three tasks.

1. We analyze the fidelity of pluri-ego-sampling, the method we used to obtain the dataset presented in our first project. For reference, our method consists in taking m initial nodes, extracting their ego networks of radius n (in this case n=2), and joining all of these graphs. This was convenient for our task for several reasons: Twitter is an extremely large network; the desired tribes likely exist in very specific regions of the full graph; *Limberg, Barnes* [6] conveniently provided the names of "chieftains" for each tribe, whose twitter accounts we used as starting nodes.

2. We try to explain the broken power law degree distribution exhibited by our network. Our dataset consists of nodes for twitter accounts and edges that represent mutual follow relationships. Twitter as a whole exhibits characteristics of a hybrid information and social network as described in [2], and by collecting only mutual follow edges, we hoped to narrow the scope to a network of social relationships.

A scale free network with a power law distribution was expected, but instead we found a broken power law with two clearly distinct exponents with the hinge around d=10. In an effort to understand this, we searched through the literature to find scarce mention of broken power laws at all, and even less attempts at explaining them. However, we tested some of our own hypothesis for what types of network generation dynamics might give rise to such a distribution. We also found examples of broken power laws found in twitter mutual networks in previous works. [2] [5]

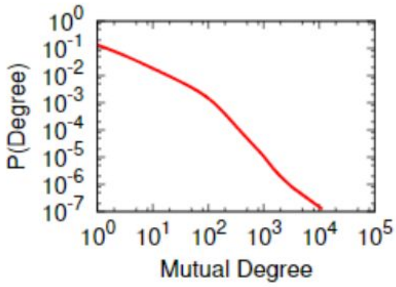
3. We attend to the original objective of the project, applying two community-finding methods - clique percolation and the Girvan-Newman algorithm - to empirically detect and identifying the communities described as *memetic tribes* in *Limberg, Barnes* [6]. We then proceed to manually and subjectively identify the tribes by inspection of the most central members according to several centrality measures.

Related Work

Although we haven't found any work with the same goals as ours, there are still several findings of interest to our analysis.

Myers et al. characterize Twitter as a hybrid social and information network. As previously mentioned, our aim has been to focus on the social structure evidenced by Twitter mutuals. Their definition of a social network rests on exhibiting characteristics observed in other social networks. "These include 1. high degree assortativity, 2. small shortest path lengths, 3. large connected components, 4. high clustering coefficients, and 5. a high degree of reciprocity. An information network, on the other hand, is a structure where the dominant interaction is the dissemination of information along edges: these are characterized by large vertex degrees,a lack of reciprocity, and large two-hop neighborhoods."[2]

In their analysis, *Myers et al.* measure and report properties of a snapshot of the mutual network from 2013 [2]. 42% of edges in the follow graph are mutual.



(c) Mutual degree (All)

Fig 1. Degree distribution of mutual network in [2].

"The presence of users with thousands of followings is indicative of "non-social" behavior. It has been well-established that individuals are only able to maintain around 150 stable social relationships at a time ."[2] Our assessments seem to have converged, as we removed these unnatural accounts with unusually high numbers of outgoing followings during the sampling process.

Average Shortest Path Length = 4.17 for mutual graph, which is consistent with other social networks[2] and somewhat close to the value found in our own network (aSPL=5.40)

"Average Clustering Coefficient = 0.23. Analysis of clustering coefficients in the Twitter mutual graph suggests that Twitter exhibits characteristics that are consistent with a social network."[2] Furthermore, this value is consistent with the our graphs value of CC = 0.26.

An important conclusion is that we're missing study of the dynamic graph of Twitter. "Hypothesis: Twitter starts off more like an information network, but evolves to behave more like a social network. [...] In particular, users accumulate followers and follow more people over time"[2]. We think studying the evolution of the graph may explain the broken power-law present in our degree distribution.

In their paper, **Kunegis et al.** [5] perform an empirical study of the preferential attachment phenomenon in temporal networks and show that on the Web, networks typically follow a non-linear preferential attachment model, contrary to the common assumption of linearity.

They find a broken power law in the outdegree distribution of a Twitter interaction graph. This is not the same as our mutual follower graph, but one would expect users that interact often to follow each other eventually, suggesting they are related.

Although this is a temporal study, it does not focus specifically on the Twitter graph, so we may still be missing some insight into the reason for the broken power law. The non-linearity of preferential attachment doesn't seem to be the reason, as "In particular, we find that the majority (70%) of the studied online networks fall into the sublinear category, having < 1." Facebook has a non-linear preferential attachment exponent beta ~0.5 and a power aw alpha of ~1.5. However, the Twitter interaction network is very close to linear preferential attachment (beta ~1, alpha ~2.5).

We expected the twitter mutuals graph to share similarities with facebook's since they involve mutual "follow" relationships. The reason here may be that the frequent interaction graph is not sufficiently similar to the mutuals graph, or that for some reason facebook friend networks are not analogous to twitter's.

Ego Sampling

Due to our goals of finding the memetic tribes described in *Limberg, Barnes, Memetic Tribes in Culture 2.0* we chose to sample the network in an unconventional way. We found the accounts of 70 of the reported "chieftains" of each tribe and sampled their ego-networks at radius 2. The goal was to raise the probability that the sample we'd be taking would more promptly include members of the tribes, while at the same time, the act of joining 70 different ego networks would dilute the bias associated with having 1 central point.

In this section we evaluate our sampling method by applying it to different scale-free networks and comparing their properties with the obtained samples.

Test hypothesis: *The ego sampler devised for this project is effective if the properties of a sample taken from a scale-free graph are similar to the properties of the sampled graph.*

To test this hypothesis we generated many scale-free graphs and sampled them. Knowing the properties of the generated graph we have a baseline for what the sampled graph should look like. The results were discouraging however, the following table shows that most attributes differ greatly. This table is obtained by calculating the average of values yielded by 10 tests. The ego sampler was ran until roughly 7% of the nodes in the original graph were explored.

	Barabasi-Albert Graph	Ego Sample
Nodes	10000	692
Edges	9999	677
Power Law Alpha	3.163	2.23 ~ 9.7 (Values show drastic variance)
Degree Assortativity Coefficient	0.0542182	0.459162
Density	0.0002	0.00290

Most of these discrepancies could be attributed to the small scale nature of the tests ran, as 10000 nodes is not comparable to the number of users in twitter. Aside from that, these discrepancies are related to the inherent structure of the sampling and therefore couldn't be eliminated by variations of the sampling algorithm or by some form of data processing we could find. With these doubts in mind we couldn't reach a solid conclusion about the effectiveness of the sampling algorithm.

We set out to test a weaker **hypothesis**: *The ego sampler devised for this project is effective if the communities found in the original graph are preserved in the sampling process.*

This simple premise makes our Twitter sample valid for the purposes of finding communities, which is exactly what we are trying to achieve. To test this hypothesis we ran a simple test: Generate a graph and run a community finding algorithm on it (in this case, Girvan-Newman) finding k communities. Then, sample the graph and run the exact same algorithm on it, finding another k communities. With these two k communities found, we attempt to match them 1 by 1 with a greedy matching algorithm based on how many nodes the sets share, which should be enough, seeing as the sets are (ideally) similar. Having a match between communities, each sampled node is tested to see if the community it is inserted in corresponds to the matching community in the original graph. We call the percentage of correctly labeled nodes the **community preservation score.** The values obtained for this score were astonishingly high. When sampling Barabasi-Albert graphs with 1.000 nodes a sample with 7% of the nodes achieves an average of 88.6% community preservation score across 10 tests. Sampling 3 Barabasi-Albert graphs with 10.000 nodes each yielded a score of 92.5%. All these tests were trying to find k=10 communities, meaning the expected result of a random labeler would be ~10%. This leads us to two very useful conclusions about the sampling algorithm:

1. This sample maintains the structure of the communities present in the original graph well enough to make them recognizable by the Girvan-Newman algorithm.
2. A small sample proves to be representative of the entire graph. The original graph had various communities spread out throughout its entire range. The sample proved to be able to probe this huge range, finding nodes from all these communities.

In the following figure this can show these 2 properties. The sample, despite being very little, spans across most of the graph, retaining a lot of data.

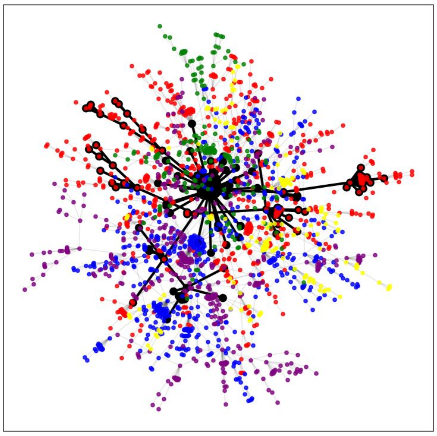


Fig: Generated Scale-free graph with 1000 nodes, and a sample with 10% its size with its edges represented..

Broken Power-law Degree Distribution

By analysing its degree distribution in our first project, the network seemed to follow two distinct exponents, this is sometimes called a *broken power-law*.

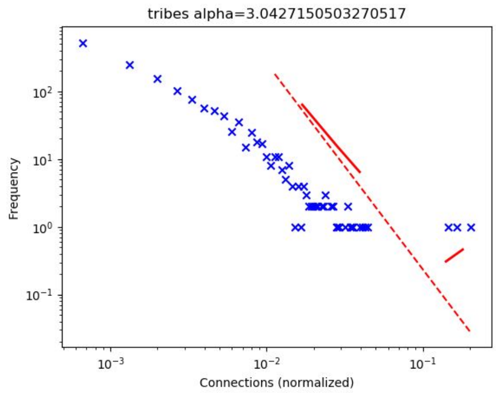


Fig: Degree distribution of a sample of 1654 nodes.

Plotting the degree distribution of a larger sample reveals what looks like a truncated power law distribution, which suggests the existence of a cutoff. In an attempt to better understand this network, we consulted two works ([2], [5]) that have also found similar distributions in mutual twitter networks.

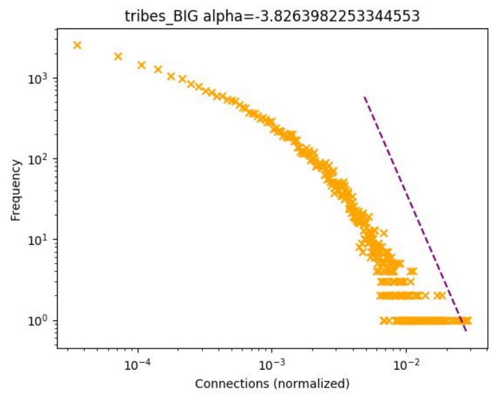


Fig: Degree distribution of a sample of 12650 nodes.

	PL	**Exp.**	**Logn.**	**Trc. PL**
p	0.004	0.996	0.253	0.039

Table: P-values for different distributions fitted using the *powerlaw* Python package.

In order: Power Law, Exponential, Lognormal, Truncated Power Law.

Fitting different distributions to our degree distribution, we find that the best fit is a Truncated Power Law - that is, a power law with an exponential cutoff - whose parameters are alpha = 3.23 , lambda = 55.95 (the exponent of the cutoff term).

	Radius 3 network
Nodes	28141
Edges	471016
Clustering Coefficient	0.22653
Degree Assortativity Coefficient	0.01923
Density	0.00104

Table: Properties of our biggest sample of the twitter mutual network.

We did not expect our degree distribution to be a perfect power law, as the amount of mutuals a user has on twitter is usually very much **cut off for high degrees** (a sort of Dunbar number number for reciprocal digital relationships). We also expected there to be some **positive assortativity** - which doesn't seem to be the case very much - and some form of **initial attractiveness** related to the fact that most new twitter accounts already know a few people on twitter who will gladly follow them back, irrespective of degree.

Interestingly, Barabasi observes in [10]: "*High degree cutoffs:* If preferential attachment is sublinear, ***the degree distribution follows a stretched exponential, or a power law with an exponential cutoff***." Which consistent both with our measurements and with the observation of sublinear preferential attachment in **Kunegis et al.** [5].

Another factor that is also not out of the question is the notion of **tag-based preferential attachment**, which we test in its linear form, and with single tags per node, but don't run multi-tag or non-linear experiments. We modified a Barabasi-Albert generator to use tag-based preferential attachment. The resulting distributions seemed like regular power laws. If there was a break, it was never very pronounced.

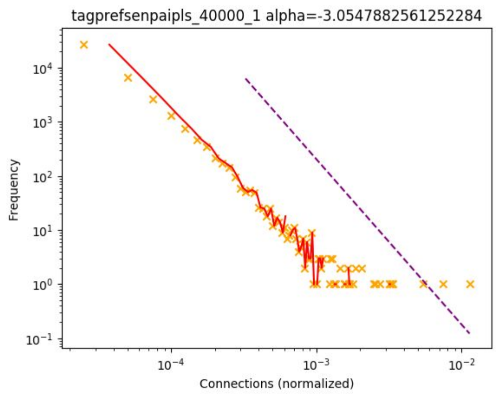


Fig: Power Law generated from tag-based BA preferential attachment.

We conclude that the best explanation for our degree distribution is that the sublinear nature of online networks is being expressed in twitter's network of mutual followers, giving rise to an exponential high-degree cutoff.

Community Finding

We applied two different Community Finding algorithms: Clique Percolation and Girvan-Newman's algorithm. After finding the communities for each algorithm, different centrality measures - the choice of which was explained in our previous paper - were applied to determine the most relevant accounts within each community.

The centrality measures applied to the nodes were:

- **Degree Centrality** - based on the number of relationships each node has (the most neighbors the node has, the more important it will be);
- **Katz Centrality** - based on how important the node's neighbors are (the more important the node's neighbors, the more important it will be too);
- **Betweenness Centrality** - based on the number of times the node appears in the path between two pairs of nodes (the more paths appear, the more important the node will be).

After this we try to identify each of these communities, an analysis was also made, calculating their parameters and comparing them. We also find that chieftains - our starting points for the sampling - aren't even the most central nodes in most cases.

For Clique Percolation, and several communities were obtained for the different values of k, however, we only considered the communities for k = 3, since the communities for the other values of k are subsets of the communities of k = 3.

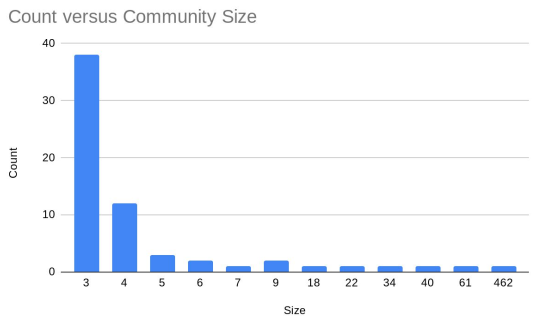


Fig: Clique Percolation's community size distribution for k=3.

Given the sizes of the communities obtained, we chose to analyze the six largest communities

We applied the Girvan-Newman algorithm successively to the dataset until we reached several communities, after which it would be useless to continue this process. This process allowed us to find ten communities, which are in the table below.

Community Analysis

Clique Percolation

The following tables include subjectively attributed labels that will be explained in the next section.

ID	Label	Size
1	Antifa/Anarchists	462
2	Post-Rationalists	61
3	Integral Theorists	40
4	Accelerationist-Right	34
5	Brickers (LEGO Culture War)	22
6	Street Epistemologists	18

Table: Subjective labels and size for the 6 communities found by the Clique Percolation Algorithm

Community	Avg. Degree	Avg. Path Length	Avg. Clustering
1	9.9	2.7	0.5
2	7.0	2.2	0.4
3	10.6	1.8	0.6
4	6.1	2.3	0.5
5	5.0	1.8	0.7
6	10.8	1.4	0.8

Table: Network properties of the 6 communities found by the Clique Percolation Algorithm.

To subjectively label our resulting communities, we ranked their nodes by centrality and observed their live twitter profiles. Communities 2, 3, 6 correspond cleanly to communities presented in *Limberg, Barnes* [6], while 1 - which is also the largest by a large margin - seems to blend the Antifa and Anarchist communities, suggesting closeness between them. We were pleased to note that our algorithm found two clear communities that were not described in the *Memetic Tribes[6]* but are nonetheless very interesting: “Brickers” are a cluster of LEGO-themed ideologically charged accounts, the new front of the culture war[7]; the Accelerationist-Right, a community marked by the ironical demonstration of anti-democratic ideas. These last two seem to be less closely connected or less popular communities overall as suggested by their lower average degree.

Average Path Length is generally small (around 2), the Average Degree and Average Clustering have overall high values. The use of this algorithm has the disadvantage that it is harder to find more communities, the larger the k the harder it is.

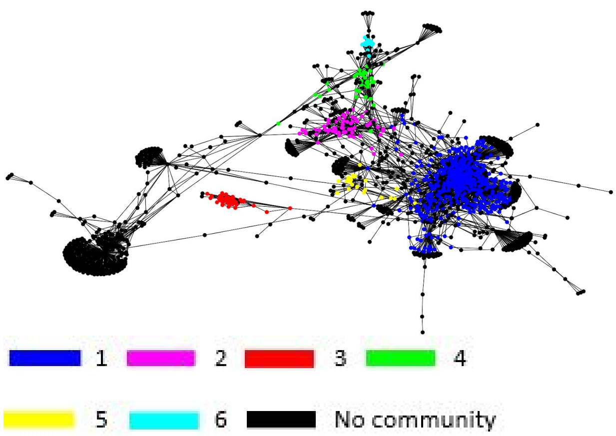


Fig: Visualization of the 6 communities found by the Clique Percolation Algorithm. The nodes in black are not present in any community.

Girvan-Newman

Id	label	Size
1	QAnon / Anti Sex Trafficking	269
2	Alt-Lite/Tea-Party/Alt-Right	69
3	Integral Theorists	44
4	Post-Rationalists	149
5	Neoreactionaries (NRx)	106
6	“Dirtbag-Left”	15
7	Street Epistemologists	93
8	SJAs / Sci-fi writers	71
9	Antifa/Anarchists	625
10	DSA/#BLM/Left NY politicians	57

Table: Subjective labels and size for the 10 communities found by the Girvan-Newman Algorithm.

The next figure represents the communities found in the graph using the Girvan-Newman algorithm, each community is represented by a different color.

Community	Average Degree	Average Path Length	Average Clustering Coefficient
1	2.8	2.2	0.1
2	2.7	2.3	0.3
3	9.8	2.0	0.5
4	4.5	3.2	0.2
5	5.4	3.3	0.3
6	2.3	2.3	0.0
7	2.9	3.2	0.2
8	2.5	3.1	0.2

9	7.9	2.8	0.3
10	3.1	2.5	0.4

Table: Network properties for the 10 communities found by the Girvan-Newman Algorithm.

Out of the 10 communities found, 3,4,6,7 form a 1 to 1 correspondence with the list in *Memetic Tribe[6]*. 2, 9, 10 seem to aggregate several closely tied tribes, respectively american right-wing politicians, Antifa and Anarchists once more the largest cluster, and american left-wing politicians - the only three possible political philosophies. QAnon, another of the predicted tribes, and alleged far-right conspiracy is surprisingly tied to anti-sex-trafficking activists. In turn, the SJA tribe seems to puzzlingly be related to many Nebula award-winning writers.

The only unexpected community was one of Neoreactionary tendencies, an ideology also referred to as the "Dark Enlightenment", which counted with the membership of Alt-Right chieftain Richard Spencer.

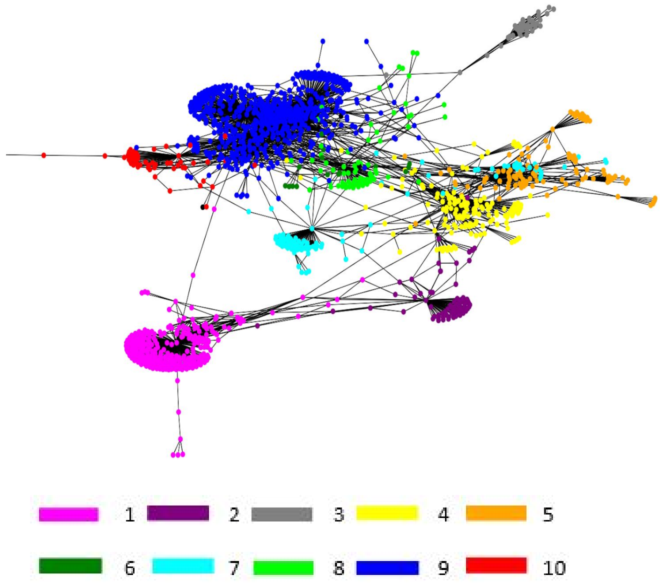


Fig: Visualization of the 10 communities found by the Girvan-Newman Algorithm.

Looking at the Average Degree, we find that there are communities that have higher values than the others (3,4,5,9), the higher variance in average degrees can be attributed to GN's higher tolerance for membership in a cluster when compared to clique percolation. Average Clustering is lower overall for the same reason.

Average Path Length is generally higher than with the first clustering algorithm. We then conclude that it is more difficult for a node to reach another within these communities and nodes are less likely to know each other.

Girvan-Newman makes it easier to find communities in the graph, but has the disadvantage that communities are not as compact.

Conclusion

We consider our project a success, with 3 goals achieved:

- The evaluation of a sampling method which proved to maintain community structure very well despite not preserving most properties of the original network,

- The analysis and explanation of an odd degree distribution, which we find is most likely a truncated power law caused by a Dunbar number of mutual digital relationships.
- The empirical confirmation of the existence of many of the memetic tribes presented in *Memetic Tribes in Culture War 2.0[6]*, as well as the amusing uncovering of the ideologically charged LEGO accounts of the Bricker community.

References

1. Smith, Jeffrey A., "Global Network Inference from Ego Network Samples: Testing a Simulation Approach" (2015). Sociology Department, Faculty Publications. 258.
2. Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin, “Information Network or Social Network? The Structure of the Twitter Follow Graph” (2014). WWW’14 Companion,April 7–11, Seoul, Korea.ACM 978-1-4503-2745-9/14/04
3. Mark Needham, Finding influencers and communities in the Graph Community (2019)
4. Shaham Farooq, Generating A Twitter Ego-Network & Detecting Communities (2018)
5. Jérôme Kunegis, Marcel Blattner, Christine Moser, “Preferential Attachment in Online Networks:Measurement and Explanations” (2013)
6. Limberg, Barnes, The Memetic Tribes Of Culture War 2.0 (2018)
7. Charlie Nash, Lego quote accounts are the new front of the culture war (2019)
8. Waclaw, Bartek & Sokolov, I. . Finite-size effects in Barabasi-Albert growing networks. Physical review. E, Statistical, nonlinear, and soft matter physics. (2007) 75. 056114. 10.1103/PhysRevE.75.056114.
9. Clauset, Aaron (2009). "POWER-LAW DISTRIBUTIONS IN EMPIRICAL DATA". SIAM Review.
10. Albert-László Barabási, Love is All You Need - Clauset's fruitless search for scale-free networks (2018)
11.