# CS 224W Final Project
# Modeling the growth of Bump

Stanislav Moreinis

Dec. 8, 2010

## 1 Introduction

The sheer explosion of mobile devices, as well as applications designed for the mobile platform, have created many usage networks which have spread throughout already existing and previously analyzed social networks. One such application is Bump, available to both iOS and Android users, which allows its users to exchange information over the Internet (such as contact cards, Facebook friend requests, or even money via PayPal) by physically bumping their phones together. Because of this combination of an online viral phenomenon with a physically constrained interaction model, an observed network of a Bump user's contacts would have effects from both virtual as well as real-world social networks. To better understand what effects dictate the current structure of the network, we will try to replicate its primary features with a scale-free synthetic model.

## 2 Network Characteristics

To gain more insight into the structure of the Bump network (and to have some quantifiable method of comparing how well a model fits the underlying data), we have chosen a set of graph properties which represent especially telling aspects of the network structure. The most obvious feature is the node degree distribution, particularly the number of leaves, as well as the average and maximum degree. While the distribution itself can tell us how well the network obeys the power law, the latter 3 statistics allow us to quantify a similarity between two different networks (such as a real and a modeled one). We can also calculate the maximum likelihood estimate of the power-law exponential coefficient $\gamma$ if we assume that the degree distribution of the network follows a power law distribution. Another important property of a connected network is the shortest-path length distribution over the network, as well as the largest shortest-path between connected nodes (i.e. the network diameter). We can also calculate the assortative coefficient, which represents the correlation between a node's degree and its nearest-neighbors' average degree. If the assortative coefficient is negative, it means high-degree nodes prefer to connect to low-degree nodes (as is the case in the Internet infrastructure), and its magnitude reflects the strength of the correlation. To further explore the connection patterns of high-degree nodes, we calculate a statistic called the rich-club ratio for some degree $k$, which is the ratio of edges between nodes of degree greater than $k$ to the total number of possible edges between these nodes (which is $n(n-1)$, given $n$ nodes with degree greater than $k$). If the rich-club ratio is 1.0 for some $k$, then all nodes of degree greater than $k$ connected, and we call this a rich-club clique. Finally, we also calculate the average triangle coefficient for a node, where the triangle coefficient counts the number of triangles a node is a part of.

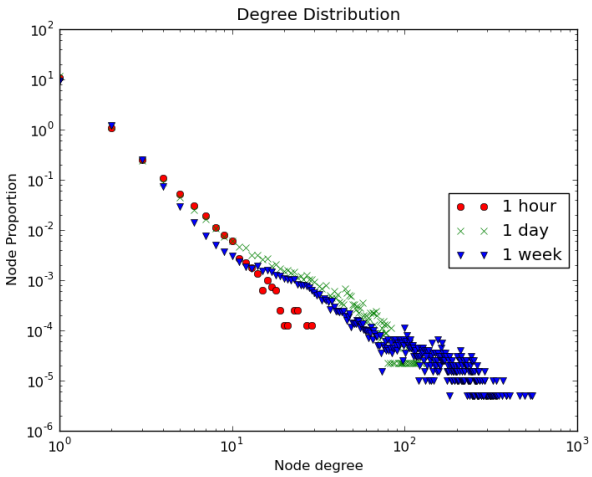
## 3 Actual Network

If we wish to approximate the structure of our underlying network using the statistics mentioned above, then clearly our first step is to determine what type of characteristics our desired network model should have. To create our actual network, we add an edge between users A and B if they have been paired by Bump at least once. Our graph is undirected (since the matching process is symmetrical), and disregards the frequency with which users communicate. To observe how this network changes as we consider a larger time interval, we have created such networks over periods of 1 hour, 1 day, and 1 week in the month of October. To account for variations in the usage, the hourly networks were averaged over the span of a day, the daily networks were averaged over the course of a week, and the weekly networks were averaged over the span of one month.

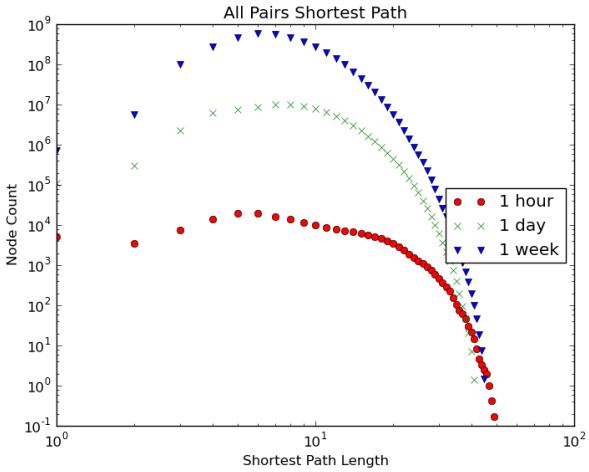
| | 1 hour | 1 day | 1 week |
|---|---|---|---|
| Nodes | 4188.375 | 85724.286 | 533844.500 |
| Leaves | 3668.792 | 74921.286 | 453352.500 |
| Edges | 2561.208 | 55958.000 | 362962.000 |
| Average Degree | 1.207 | 1.301 | 1.358 |
| Maximum Degree | 14.042 | 116.714 | 510.750 |
| Assortative Coeff | 0.481 | 0.399 | 0.378 |
| Graph Diameter | 20.750 | 38.000 | 40.500 |
| Avg. Triangle | 0.029 | 0.204 | 1.175 |
| Rich Clique Degree | 3.04 | 50.714 | 249.000 |
| MLE $\gamma$ estimate | -11.049 | -9.457 | -8.453 |

Comparing the structure of the 3 different granularities, we can see that the number of edges increases proportional to the number of nodes, and the average degree remains fairly constant, much like the MLE $\gamma$ estimate (which remains a large negative number despite the change from the daily to weekly networks). The average triangle coefficient is higher for models larger than 1 hour (since the triad effect is less likely to happen in the span of an hour as it is in the span of a day or a week). The rich clique degree (proportional to the maximum degree) is also fairly similar, and assortative coefficient suggests all 3 networks have a slightly assortative structure.

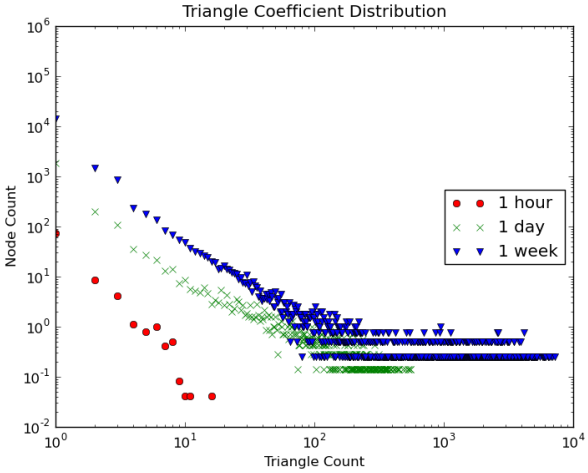
Although we have shown that the networks are (fairly) similar at differing granularity levels (increasing predictably as a larger time period is considered), we have also seen that compared to a typical network in which the power law effect is observed, our considered networks have both a disproportionately high number of leaves, as well as a very high value of $\gamma$ (typical values for it being between -2 and -3). The large number of leaves can be explained by Bump's dynamic - there is a large base of the user population which only bumps a small social circle, which starts overlapping with others only once it has grown beyond some size. The largest connected component in our network only accounts for roughly 11% of all of the users (and the 2nd largest has a much smaller size), so the majority of the nodes do not exhibit many scale-free growth effects since they exist in isolated tiny connected components. A potential cause of such sparse connectivity of the network is the physical constraints for interaction, which limits the number of potential people one can match. We can see the network statistics, as well as the distribution of the sizes of the connected components across the three different graphs below:



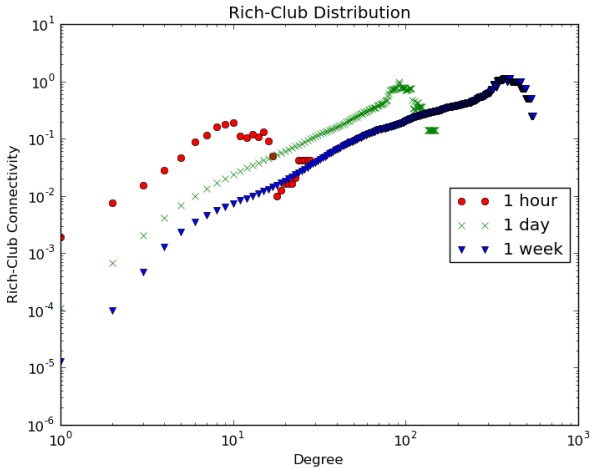

(a) Node Degree




(b) Shortest Path

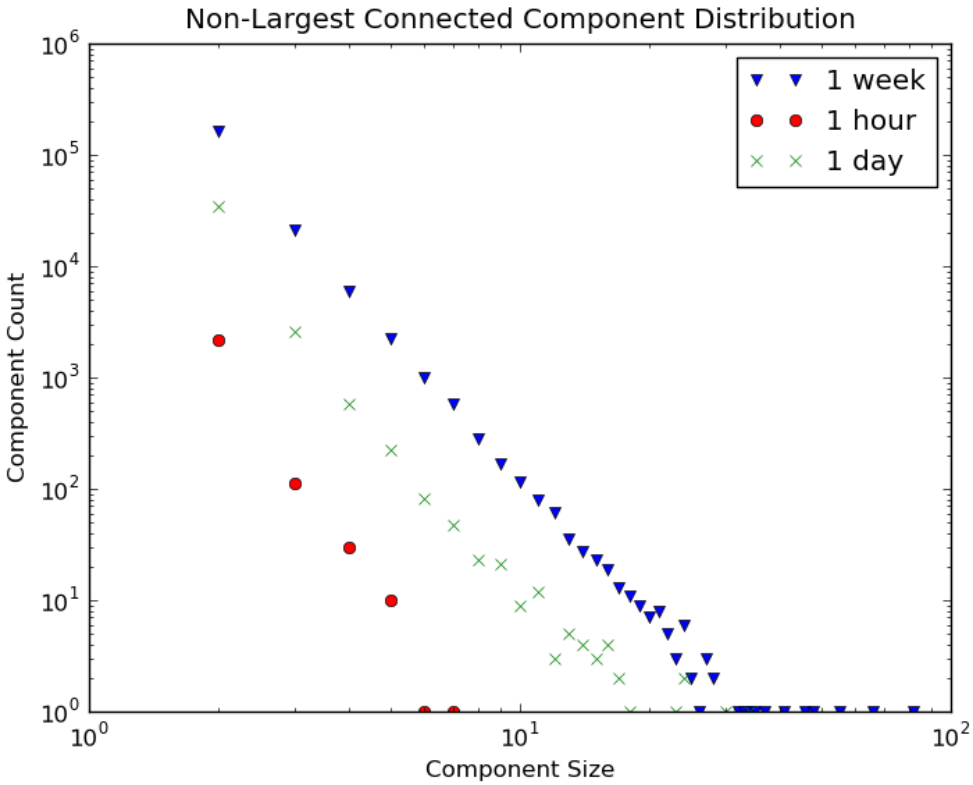



(c) Triangle Coefficient




(d) Rich Club

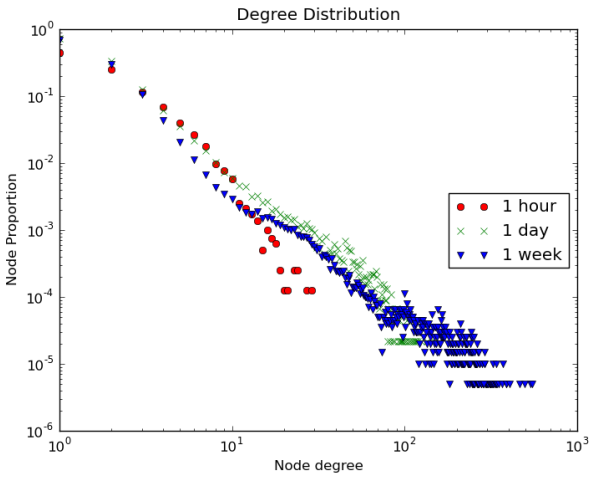
Figure 1: Network Characteristics



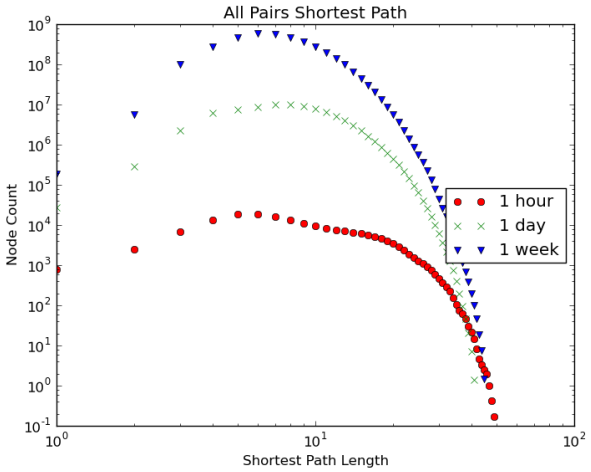

If we consider only the largest component for each of our graphs, we obtain the following statistics:

| | 1 hour | 1 day | 1 week |
|---|---|---|---|
| Nodes | 335.375 | 9074.286 | 60639.750 |
| Leaves | 150.5 | 4669.143 | 34218.000 |
| Edges | 413.75 | 13720.429 | 92865.500 |
| Average Degree | 2.318 | 3.012 | 3.057 |
| Maximum Degree | 13.292 | 116.714 | 510.750 |
| Assortative Coeff | 2.098 | -0.577 | -0.136 |
| Graph Diameter | 20.667 | 38.000 | 40.500 |
| Avg. Triangle | 0.141 | 1.769 | 9.871 |
| Rich Clique Degree | 3.333 | 50.714 | 249.000 |
| MLE $\gamma$ estimate | -2.679 | -2.717 | -3.056 |

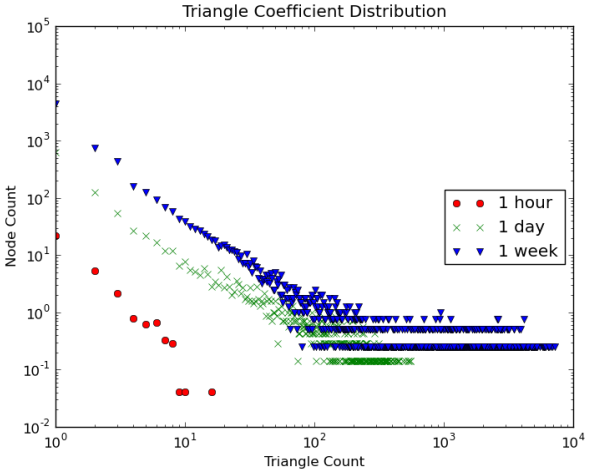
The first observation we can make is that the general structure of the largest connected component (even though it does not contain the majority of the nodes) mirrors the structure of the network as a whole in many aspects. The leaves now make up a much smaller proportion of the graph (being reduced by about 30%), and since so many lower degree nodes have been removed, we can see that the average degree has increased to about 3 edges per node. As another result of this modified degree distribution, our maximum likelihood $\gamma$ estimate is now 3.056 for a weekly network, which is a much more typical number for scale free networks. The average triangle coefficient (especially when considered over the course of one week) is much higher, which is a reasonable result considering that all of the nodes in the connected component are much more likely to form triangles between each other since they have some finite degree of separation



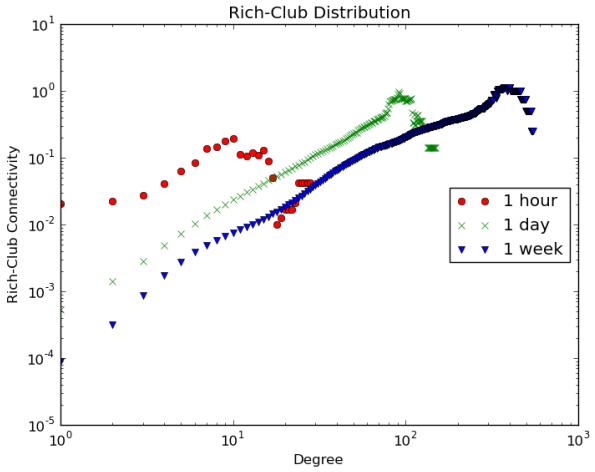

(a) Node Degree




(b) Shortest Path




(c) Triangle Coefficient




(d) Rich Club

Figure 2: Largest Component Characteristics

from each other. Finally, we can also see that the assortative coefficient has changed from being slightly positive to slightly negative for the connected component. Although it looks as though this would signal a different network structure, this change is also caused by the removal of low degree nodes. In the largest connected component, the average neighbor degree of nodes with degree less than 5 is much higher than that in the network as a whole, since those present in the connected component are more likely to be there by having an edge to a high degree node (and if they only had a few edges to low degree nodes, they would most likely be in a much smaller connected component). Because the lower degree nodes also account for a large proportion of nodes in the graph, their new average neighbor degree causes the assortative coefficient to change sign.

# 4 Barabási–Albert Model

## 4.1 Theoretical Foundations

Since we have seen that our observed network exhibits scale-free properties, we can use some scale-free network generation algorithms to try to model it. One of the first models that attempted to model an emerging power law effect was the BA model of preferential attachment. For any given user that is introduced to Bump by a current user, it seems reasonable that those nodes with a high degree count (i.e. those nodes which are already well-connected) should be more likely to have introduced the user to the service compared to nodes with a low degree count (which exhibit a low amount of activity in our given network). This concept is called preferential attachment, and it is the cornerstone of the Barabási–Albert model (which assumes a linear relationship between a node's degree count and its likelihood of being connected to a new node), as the probability of the new node being connected to a node $i$ with degree $k_i$ is $p_i = \frac{k_i}{\sum_j k_j}$. The BA model is then generated by starting with an initial core of $m_0$ nodes, and at each step creating a new node with $m$ edges to already existing nodes. The result is an almost certain creation of a few highly-connected nodes (reflecting the phenomenon of few celebrities being highly-connected in social networks or few sites such as Wikipedia being highly-linked on the Internet), and results in a degree distribution that is more like the one observed in social networks or the Internet in that it obeys the power law [1].

## 4.2 Simulations

While the BA model replicates the scale-free nature of networks, it is not flexible enough to model the degree distribution of our observed network. For values of $m > 1$, the resulting models do not have any leaves, since each created node already has several edges. For $m = 1$, the resulting model was too sparse and resulted in high-degree nodes that did not have a sufficient number of edges.

# 5 Positive-Feedback Preference Model

## 5.1 Theory

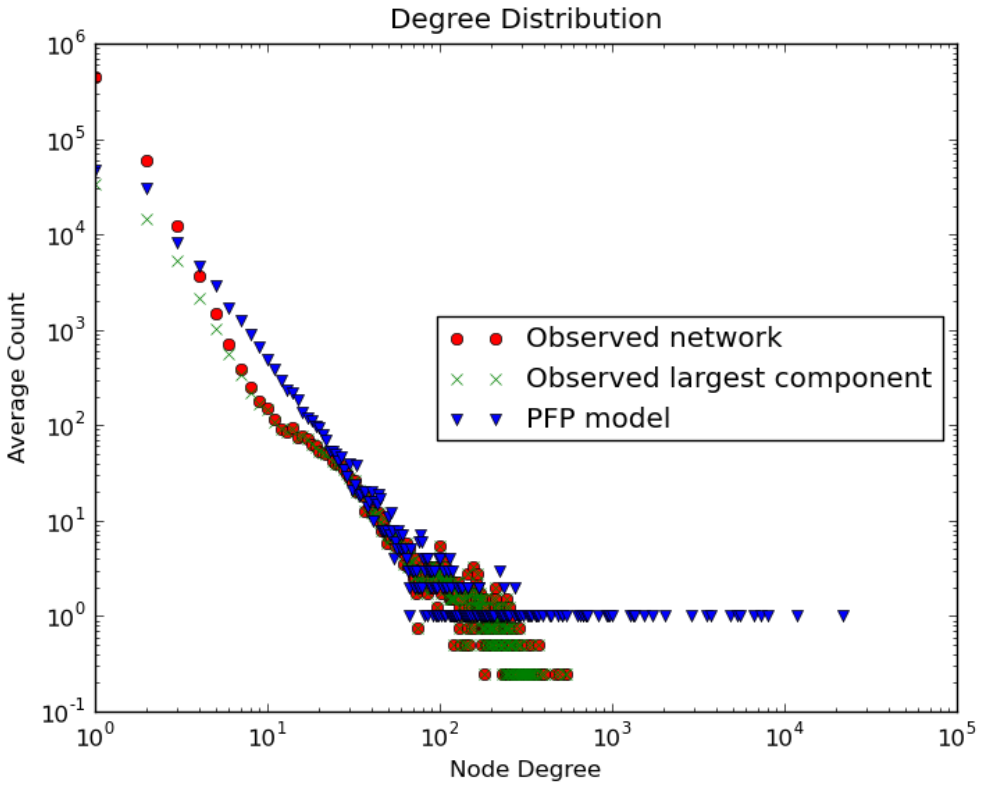
To help address the rigid structure of the simulated BA network, as well as the rate with which high-degree nodes become more and more preferred, we introduce the positive-feedback preference model. While the BA model simply adds a node with $m$ random edges at each step (which are more heavily weighted towards nodes of higher degree), the PFP model considers two distinct

possibilities of a new node joining the network. With probability $p$, we create an edge between the new node and some existing node $A$, as well as two random edges from $A$ to existing nodes. With probability $1-p$, we create edges between the new node and two existing nodes $A, B$, and then create a random edge from either $A$ or $B$ to another existing node. When choosing nodes as endpoints of edges, we also treat nodes with higher degree more favorably, but unlike the linear relationship the BA model suggests, the probability of choosing a node $i$ with degree $k_i$ is $p(i) = \frac{k_i^{1+\delta \log_{10} k_i}}{\sum_j k_j^{1+\delta \log_{10} k_j}}$ for some constant $\delta$. Compared to the BA model, this not only makes nodes with higher degree more favorable to gain edges, but also makes nodes become disproportionately more favored as their degree becomes higher and higher. The PFP model has been shown to emulate models of the Internet rather admirably [2], so it should do well to at the very least model the purely virtual aspects of the way the Bump network has grown and evolved.

## 5.2 Simulations

## 5.3 Degree Distribution

Since this model is intended to primarily model the degree distribution of scale-free networks, we used the degree distribution of the simulated model to estimate its accuracy in reflecting the behavior in the observed network. Using the cited value of $\delta = 0.048$, we have found that low values of $p$ (e.g. 0.1) give us degree distributions that are mostly comparable to the observed one. The most striking differences are the extremely low number of leaves (as each new node is almost certain to be created with 2 edges to existing nodes), as well as the few nodes with exceptionally high degree. To alleviate both of these issues, we have introduced another step into the PFP model to more closely model the behaviour of new Bump users. Since new leaves are only created with probability $p$, which also causes existing nodes to gain an additional 2 edges, we decided to decouple these two processes by introducing another probability threshold $p'$. With probability $p'$ (0.6 in our experiments), we create a leaf, which causes two new edges for an existing node with probability $p$ (e.g. 0.9), and introduces no other edges into the network otherwise. This helps us account for the larger amount of leaves observed in the Bump network without having to saturate the highly connected core of the network. We can see how this variation of the PFP model performs below:

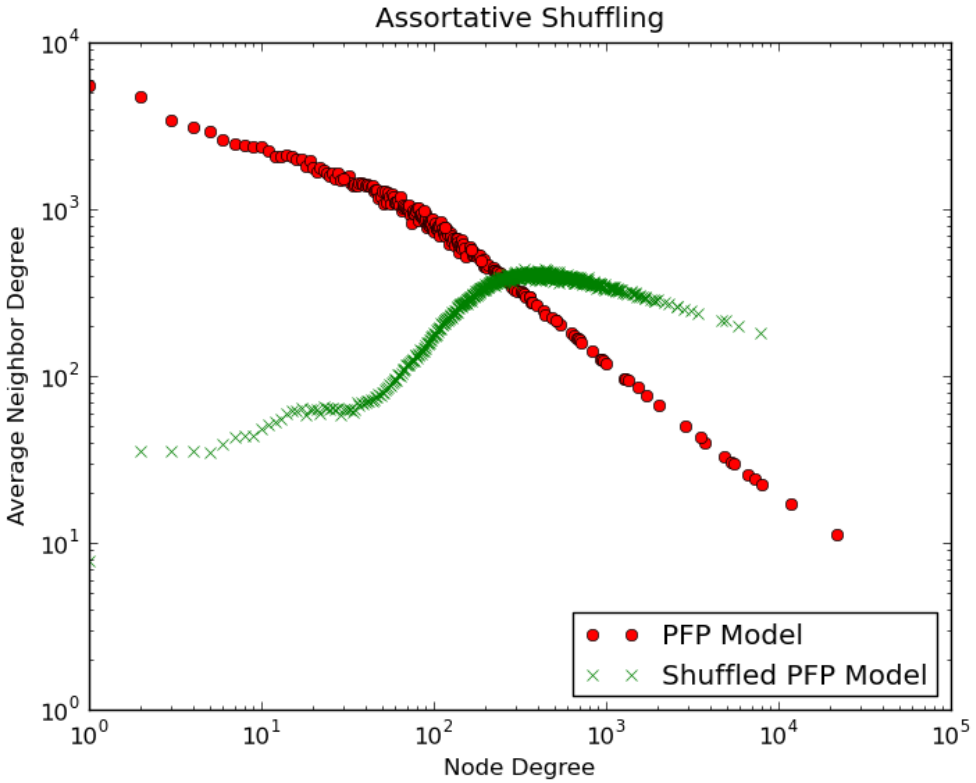



The resulting degree distribution is fairly comparable to the observed one, especially for nodes with degree larger than 10. One of the differences we can see is that there are more nodes of degrees 1 to 3 observed than expected, and at the same time less nodes of degrees between 4 and 10 observed than predicted. I believe that this can be accounted for by looking at a larger timespan for our model, as many of the nodes which we have recorded as having degrees 1 to 3 would very likely be seen matching several other users, resulting in the straight line predicted by our model. While the general distribution of highly connected nodes is fairly well predicted, there clearly are very high-degree nodes which are predicted by the PFP model, but were not observed in our network. Although observing our high-degree users over a larger time window would be likely to increase their degree (much as the low-degree users), I believe that the physically constrained interaction model of Bump plays a part in placing an upper bound on a node's number of neighbors. Because the application has yet to achieve ubiquity, and not everyone (exposed to it or not) has decided to become an adopter, the high-degree nodes are very likely to simply saturate their area and match everyone around them who uses or will use Bump, preventing any further increase in degree that would be expected by the PFP model.

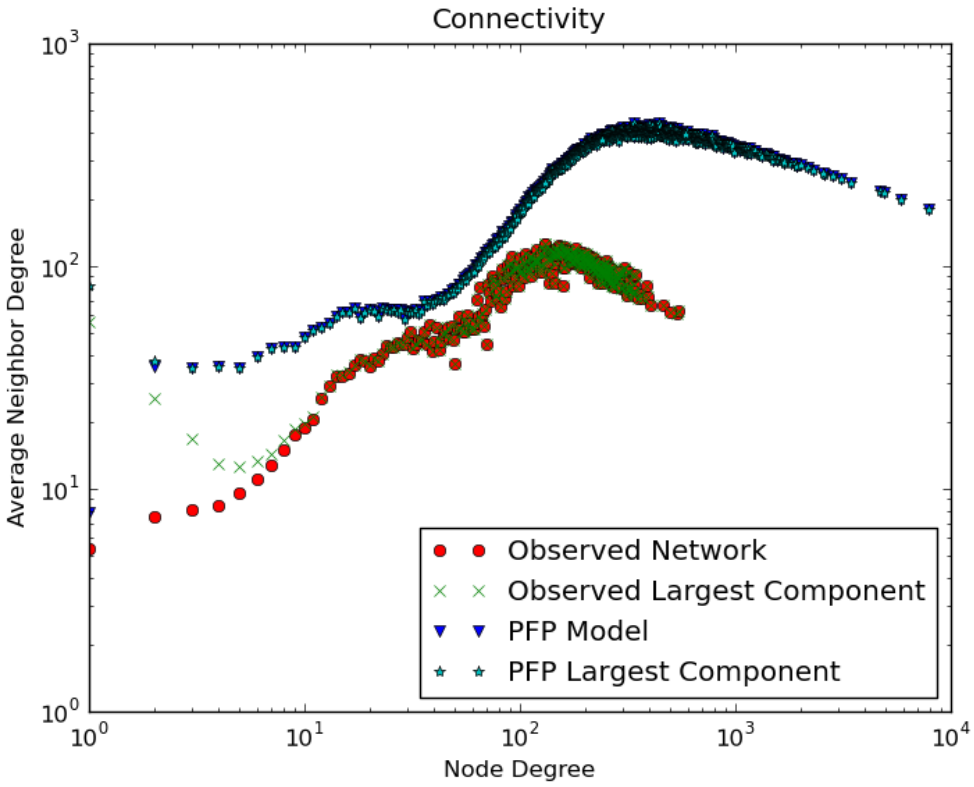
## 5.4 Node Connectivity

As the PFP model was designed to model the structure of the Internet (with not only the scale free growth, but particularly having the rich get disproportionately richer), it also displayed dissasortative aspects also displayed by the Internet structure, where low degree nodes generally connect to high-degree ones and vice versa. Even though the low degree nodes in our largest observed connected component do tend to have a fairly high average degree, this effect is only

observed for degrees less than 5, after which the slight assortativity observed in our network as a whole can be seen again. Therefore, to better model the interactions between nodes, I decided to apply assortative shuffling to random pairs of edges, where with a high probability $p$, the existing edges would be removed and the two low-degree nodes and the two high-degree nodes would be instead connected together. With probability $1 - p$, the edges would be rewired randomly [3]. Since this does not change the amount of edges per node, our observed degree distribution will not change, but it will make the network exhibit the assortativity that is generally observed in social networks.




We can see that the rewiring of the edges effectively reverses the trend in our PFP model, although the average degree of low degree nodes is still fairly high. Comparing with the average neighbor degree of the observed network, we can see that although the leaves have fairly similar average neighbor degrees, other nodes with degrees less than 10 in the model network have a relatively high average neighbor degree, especially compared to nodes of higher degree in the same network. We can also see the similar problem comparing the largest connected components of the both networks. Further corrective measures to better fit the connectivity of the model to that of the actual graph will also create more and more connected components, and reduce the size of the largest one (which is currently roughly 90%, but would fall as more and more low degree nodes are connected to one another). We also see that the average neighbor degree of all nodes in the model is higher than that of the observed network, but this can be explained, much as in the last section, both by the incompleteness of the neighbor count in our observed network, as well as the saturation of geographical areas which creates an upper bound for the amount of neighbors any user might have. By accounting for these factors, our resulting connectivity distribution would be even more similar to the one we observe.




## 6 Next Steps

As mentioned before, collecting network data over a longer time window would make the conclusions to be drawn from the resulting distributions even more concrete, as we could tell whether disproportionate number of low degree nodes is a consequence of the graph structure, or an artifact of the data observed. Furthermore, removing enough high-degree nodes to account for the physical constraints involved with using Bump would not only improve the fit of our degree distribution, but would also decrease the average neighbor degree. We could then tune the assortative reshuffling to ensure a better fit of the connectivity of the synthetic model to the observed one, using the size of the largest component, as well as the sizes of the smaller ones to gauge the quality of the fit.

## References

[1] Réka Albert; Albert-László Barabási, *Emergence of scaling in random networks*. Science 286, 1999.

[2] Shi Zhou, *Why the PFP Model Reproduces the Internet?*. IEEE International Conference on Communications, 2008.

[3] R. Xulvi Brunet; I. M. Sokolov, *Reshuffling scale-free networks: From random to assortative*. Physical Review E, Vol. 70, No. 6. (2 Dec 2004), 066102.