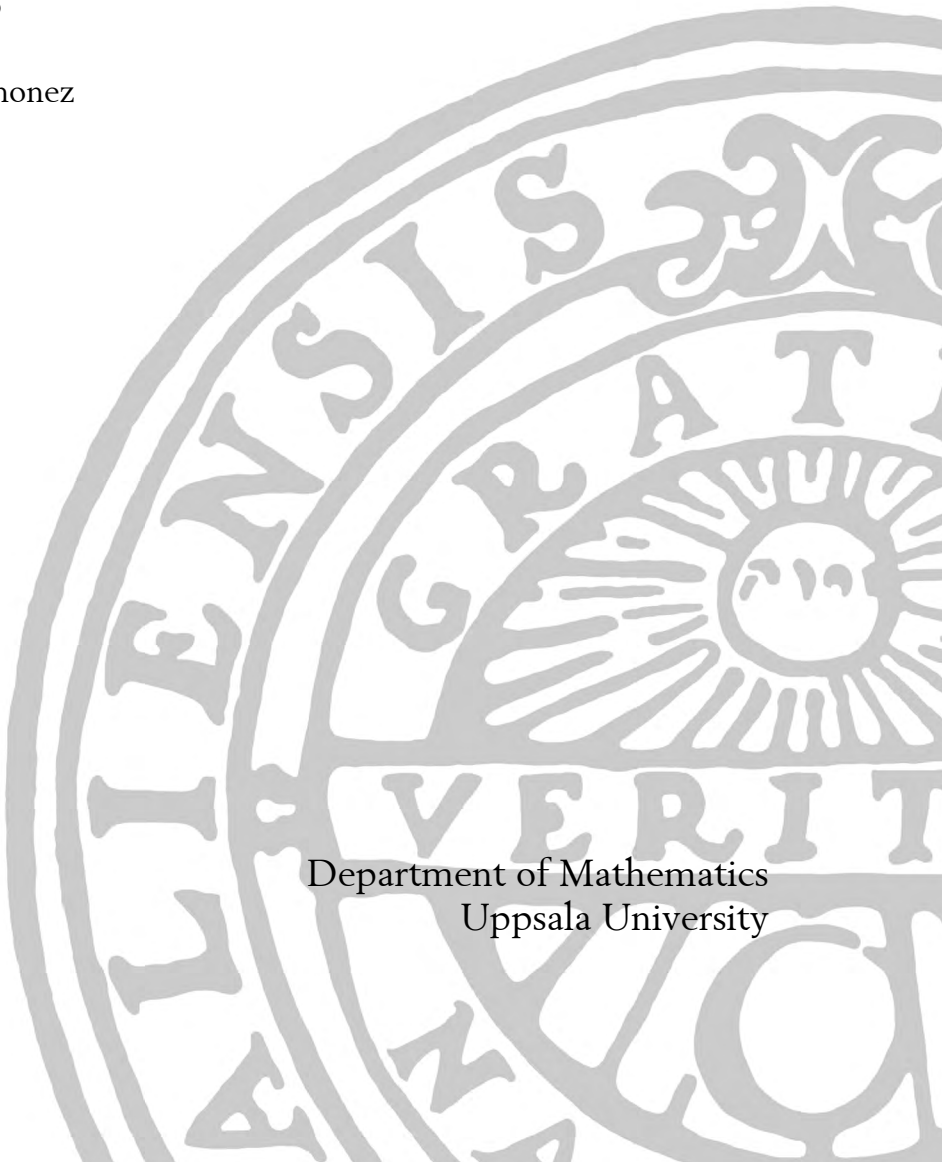
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2019:1

Screaming in to the void: Homophily in social networks

Pierre Hedström

Examensarbete i matematik, 15 hp
Handledare: David Sumpter
Examinator: Veronica Crispin Quinonez
Januari 2019



Department of Mathematics
Uppsala University

Matematiska institutionen Uppsala Universitet

Examensarbete i matematik 15 hp

Screaming in to the void: Homophily in social networks

Author
Pierre Hedström

Supervisor
David Sumpter
Alexander Szorkovszky

January 18, 2019

Abstract

This is a study set out to test if behaviour regarding tweet, retweet, statuses or a new tweet/retweet ratio can be predicted using follower count on the social media platform Twitter. This study will also test the prevalence of assortivity with regards to follower count, tweet, retweet, statuses or the tweet/retweet ratio. Tests were conducted using the Twitter API, using a network created from users that follow the user Blondinbella on Twitter. Results show possible assortivity with regards to follower count, however, where users tend to follow other users with fewer follower than themselves. This can,however, be explained by the power law behaviour of the underlying degree distribution. Results also show a correlation between in-degree and statuses count.

Contents

1 Introdution

Social media is becoming an ever growing part of our social life. According to Facebook, there was an average of 1.45 billion daily users on their platform in March 2018 [1]. Social media platforms such as Facebook, Twitter and Instagram allow users from all over the world to connect and allow us to expand our networks to include individuals we might never have met in real life with the click of a mouse or the touch of a phone screen.

This thesis had two objectives, the first was to explore the possibility that behaviour regarding tweets, retweets and statuses can be predicted using follower count. That is, a users in-degree was tested against their number of tweets, retweets and statuses to see if they was proportional in some way. A user with an high in-degree was expected to have a lot of tweets and retweets, on the other side, a user with an low in-degree was expected to have a low number of tweets and retweets.

The second objective was to explore the possibility of assortivity in the network. The characteristics tested were: number of tweets, number of retweets, number of statuses, follower count and the tweet/retweet ratio. If there is assortive mixing for any characteristic, then two nodes that are connected in the network should have equal values of the characteristic measured. If it is the case then a majority of the edges will start and end in nodes with similar values in the measured characteristic. In a network with perfect assortive mixing, the characteristics studied will be the same for the origin and end of an edge, for every edge in the network. This is highly unlikely, more reasonable is to expect that the characteristics studied will be similar, so that if we were to select any edge in a directed network, the probability that we select an edge at random that start and end with users that have similar values on that characteristic. If this is the case, the network can be said to have assortive mixing with regards to that characteristic.
However,it is not probable to expect that a network will be perfectly mixed, instead there might be a degree of mixing deemed sufficient. To decide the degree of mixing, the assortivity coefficient is used. The assortivity coefficient is described in section 2.7

A Twitter network was selected, belonging to a Swedish entrepreneur, nicknamed Blondinbella. This network consists of the users that follow Blondinbella.

Mathematicians, and sociologists alike, have studied social networks before the invention of social media platforms such as Twitter. One such study was the “Six Degrees of Separation”, which proposes that you can reach any human on the planet through six handshakes, was proposed by Frigyes Karinthy in 1929 [2]. This theory was tested by Jeffery Travers and Stanley Milgram in the, today, classic paper: ”An experimental study of the Small world problem”[3]. The study was conducted as follows:

> "An arbitrary 'target person' and a group of 'starting persons' were selected, and an attempt was made to generate an acquaintance chain from each starter to the target. Each starter was provided with a document and asked to begin moving it by mail towards the target. The document described the study, named the target and asked the recipient to become a participant by sending the document on. It was stipulated that the document could be sent only to a first-name acquaintance of the sender. The sender was urged to choose the recipient in such a way as to advice the progress of the document toward the target.
> Several items of information about the target were provided to guide each new sender in his choice of recipient. Thus, each document made its way along an acquaintance chain of infinite length, a chain which would end only when it reached the target or when someone along the way declined to participate. Certain basic information, such as age, sex and occupation, was collected for each participant."[3, p. 428].
> The result from this letter study was that, most of the letters that reached the target individual did so with a chain length of around 5. The mean value was 5.2 for all letters [3, p. 432].

If we were to observe a Small world network from a mathematical point of view, we would describe a Small world network as a network with a short path-length and high clustering[4]. If we compare a Small world network with a regular lattice, were each node is placed in the corner of a square, connecting to four other nodes, the Small world network would appear to be more chaotic, not as chaotic as a random network however, but compared to a regular lattice, the Small world network would appear quite disorganised and chaotic. This is however what gives the Small world network its strength. For with high clustering comes a shorter path length [5]. The strength in a small world network originates from the resilience of the network when one or more edges in the network is lost.

To generate a Small world network, we start with a uniform circle network, with a fixed number of nodes and edges. We then remove an edge between two nodes, and re-attach it to a new position. If this process is repeated, where probability that an edge attaches to a node is uniform over the network, until the path-length gets shorter and the clustering is increased, we end up with a Small world network [4].This generation method is called the Watts-Strogatz model. One important factor to remember is that repeated removal and reattachment can produce isolated points in the network, where two distinct parts of the network depend on a single edge to stay connected. This network is not as resilient as one that might have preceded it in the removal and reattachment process. It might be that a sweet spot can be found during the process in which the benefits from a short path-length and high clustering are maximised and

the risk of isolating parts of the network are minimized.

Not everyone was content with this model however, Albert and Barabási criticised the Watts-Strogatz model as a network model for social networks, partly because it assumes the number of nodes to be constant over the formation of the network. As a counterproposal, they instead suggest a network-model using preferential attachment. The preferential attachment also goes against the Watts-Strogatz model. In the Small world model, each node is just as likely to get a connection as any other. However, in Albert and Barabasi model of scale free networks, the preferential attachment means that not all nodes are as likely to attract new nodes. In fact, some nodes will attract lots of new edges, while others might only attract one or two. Therefore, it appears not all nodes have the same chance to get a new connection when a new node is introduced. Instead the node that is most likely to get new edges is the one that is already the most connected node in the network, therefore we get a "rich-get-richer" phenomena.

The degree function P(k) follows a power law in a scale free network $k^{-\gamma}$, and this is why a minority of the nodes in the network will have the majority of the connections[6]. This means that the logarithm of a degree distribution of a scale free network can be roughly approximated by a straight line[8]. If $P_k = Ck^{-\gamma}$ represent an scale free network, then the linear function $lnP_k = -\gamma lnk + C$[8]. γ is the exponent of the power law, and has been found by Newman to typically range between $2 < \gamma < 3$[8]. To examine a network to find out if its degree distribution is scale free or not, one method is to construct a log-log histogram, and examine to which extent the bars of the histogram follow a line, the fit must not be perfect, for small and large k values, the histogram might deviate slightly[8].

The phenomenon of assortivity has also been known and studied since before the rise of social media. Even as far back as the Greek philosophers Aristotle and Plato talked about the fact that individuals that liked each other seemed to share some kind of similarities[9]. Miller McPherson, Lynn Smith-Lovin and James M Cook studied assortivity in social networks in their review article: Birds of a Feather [9]. McPherson et al. describes different characteristic, such as age, race and religion and to which degree they appear as factors for assortivity in groups. An important fact is that groups, and members in a group, can have more than one type of relation, and that they can also have assortivity in more than one characteristic[9]. A group in which many members share the same characteristics is said to be homogenic.

One very important fact stated is that of majority and minorities. For example, in a workplace, one ethnic group might be a minority, which means that if a member of the minority group makes acquaintances at work, there is a high statistical probability that they will make acquaintances with the majority group. If a member of the majority group would to make acquaintances at

work, the probability that they will form an acquaintance with another member of the majority group is greater than the probability that they will form an acquaintance with the minority group. This is due to the distribution between the ethnic groups. Peter Blau, Terry Blum and Joseph Schwartz published a paper Heterogeneity and Intermarriage, in which they discuss the impact of distributions and group size for inter-group marriage and out of group marriage[11]. Blau et al found that there is an inverse relation between a groups relative size and the amount of inter marriage in that group[11].

2 Background

2.1 Nodes and Edges

A mathematical representation of a network can be constructed using nodes and edges. We define a node as one single point in the network, usually denoted by a dot or an intersection of edges. This node represents one entity of what the network maps, (Fig.1a). For the case of social networks, a node is one user, a person in the network. For other kinds of networks, such as train-lines, a station would be a node. The number of nodes in a network is denoted by n.

We define an edge as a relationship between two nodes. This relationship connects the nodes and is graphically represented by a line between the nodes. The number of edges in a network is denoted by m.

Edges can be either directed or undirected. An undirected edge represents a symmetric relationship (Fig.1b), where both connected nodes have the same relationship to one another. A directed edge (Fig.1c), represents an asymmetrical relationship, where one of the nodes has a different relationship to the other. One example of an undirected edges is two users on a social media platform that are following each other. For example, friends on Facebook have an undirected edge in their network, this is because both users need to befriend each other in order to be friends on Facebook, therefore the relationship is symmetric. Here the nodes are the two users and the edge represent the friendship between the users. Since the relation is symmetric, both users share the same relation in the network, the edge is undirected. If, however, only one of the users would be following the other, as is possible on the social media platform Twitter, the relationship would no longer be considered symmetric. In this case we have a directed edge, in the graphical representation an arrowhead indicates were the edge is directed to, and from.

(a) *(b)* *(c)*

Figure 1: Image "a" show two nodes as they would be visually represented. As these two nodes are represented, there is no apparent relation between them, this can be concluded by the fact that that there is an absence of an edge between them. Image "b" shows two nodes that have some kind of relation, due to the fact that they have an edge connecting them. This edge is undirected, so the relation is of a kind that must be mutual. Image "c" shows two nodes connected by a directed edge, that means that the relation between the nodes is not symmetric.

2.2 The adjacency matrix and the mixing matrix

For a network an adjacency matrix A can be constructed, where the rows and columns represent nodes in the network[8]. For an unweighed network, the matrix elements A_{ij} in the matrix can take two values, one and zero, where a one indicates an directed relation from node j to node i[8]. For an simple undirected network, the adjacency matrix is symmetrical. An example of a adjacency matrix can be found bellow:

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1j} \\ A_{21} & A_{22} & A_{23} & \dots & A_{2j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{i1} & A_{i2} & A_{i3} & \dots & A_{ij} \end{bmatrix} = A$$

Figure 2 is a graphical representation of a undirected simple network, consisting of ten nodes and eleven edges:

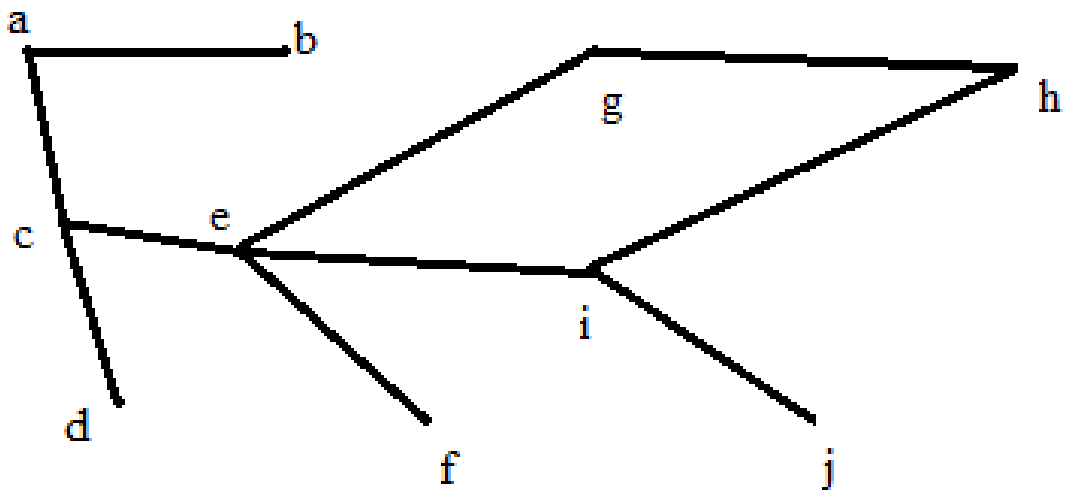


Figure 2: Network used as an example through this thesis. The network is undirected with n=10 and m=10.

This network can be represented by the adjacency matrix:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} = A^{example}$$

It is clear from this example that even small simple networks might create large adjacency matrices.

The mixing matrix e is a matrix consisting of the probability that an edge connect a node of type a with a node of type b for a characteristic i.

$$\begin{bmatrix} e_{11} & e_{12} & e_{13} & \dots & e_{1j} \\ e_{21} & e_{22} & e_{23} & \dots & e_{2j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{i1} & e_{i2} & e_{i3} & \dots & e_{ij} \end{bmatrix} = e$$

To create a mixing matrix for the example network, we first need to give every node membership to a group, lets say gender, were: a,d,f,h and i are male and b,c,e,g and j are female. Now lets define the edges to be handshakes,

since a handshake is a symmetric relation, the edges are undirected. The mixing matrix $e^{example}$ would be:

	Men	Women
Men	.1	.35
Women	.35	.2

$$\begin{bmatrix} .1 & .35 \\ .35 & .2 \end{bmatrix} = e^{example}$$

Here e_{ij} is the fraction of nodes that connects a man to a woman, a woman to a woman or a man to a man. There is 7 edges of 10 connecting women to men, however to satisfy the sum rule $1 = \sum_{ij} e_{ij}$, we divide $e_{2,1}$ and $e_{1,2}$ with 2.

2.3 Degree and degree distribution

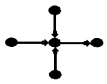


Figure 3: The node in the center has in degree 3 and out degree 1. To get the in and out degree for the node, simply count the directed edges going in and out of the node.

For an undirected network, we define the degree, k, to be the number of edges that connect the node i to other nodes in the network. For a network with only undirected edges the calculation for the degree of a node is very straight forward, one simply counts how many edges connect to a specific node.

To calculate the degree for a node i using the adjacency matrix of an undirected network with n nodes, use equation 1. This equation counts the number of values on the specific nodes row in the adjacency matrix.

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{1}$$

For example, the node e in A^{exaple} have degree $k = 4$

However, for directed networks we need to define in-degree and out-degree. This is a necessity since the relationship could be asymmetrical. An undirected edge can be seen as a special case of a directed edge in a relationship that is always symmetrical. If we use the example of train stations along a train track, a station is either connected to the network through the train track or it is not. There is no way for the station to be in an asymmetric relationship to any other station in the network.

In-degree k_i^{in} is the total amount of edges that are directed towards the specific node. The out-degree k_i^{out} is defined as the total numbers of edges that originate in the specific node and spread out towards other nodes, see figure 2.

The in- and out degree for a node i can be calculated by:

$$k_i^{in} = \sum_{j}^{n} A_{ij} \qquad k_i^{out} = \sum_{i}^{n} A_{ij} \tag{2}$$

For a network with n nodes, where the i:th node has degree k_i, we can create a degree distribution for the network. The degree distribution for the network, P(k) provides us with the probability that a random node in the network has degree k[12]. The degree distribution can be illustrated by a histogram.

The degree distribution for $A^{example}$ is illustrated using a histogram in figure 4:

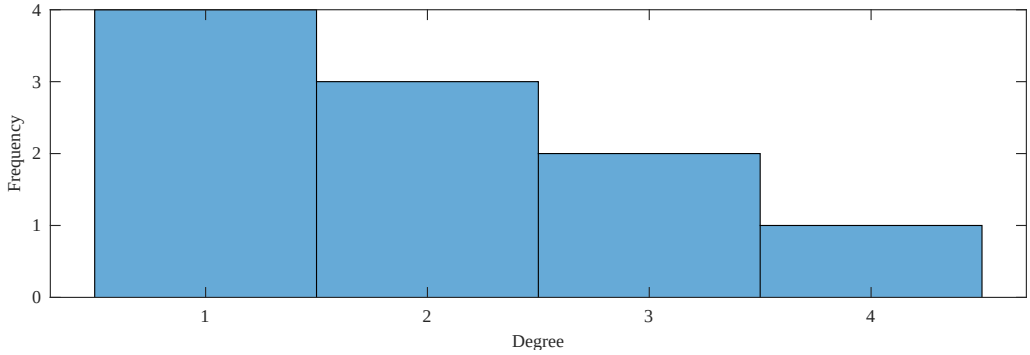


Figure 4: The histogram of $A^{example}$, note that the histogram follow the function $y = -x + 4$.

2.4 Path length

We define the geodesic path to be the shortest distance between two different nodes, where we never cross our own path, that is we never "travel" the same edge twice [5, p. 53][8]. The path length is the number of edges that we walk to get from a node i to another node j. The path length between the nodes i and j in Figure 3 is two.

The path length can be calculated by using a method called breadth-first, in which a node in the network i selected as the starting point of the measure, this node will of course have distance 0 and its neighbours, all nodes connecting to it, have distance 1. Nodes that connect to nodes with $d = 1$, that does not connect to the original node will have distance 2, and as the mapping continues out from the start-node out into the network, the count is increased by one[8]. From this follows that: for a node in the network, s, with a shortest distance d from node t, has a neighbour with a shortest distance to t that is $d - 1$[8]. Now for every node that connects to a node with distance d, that has not already

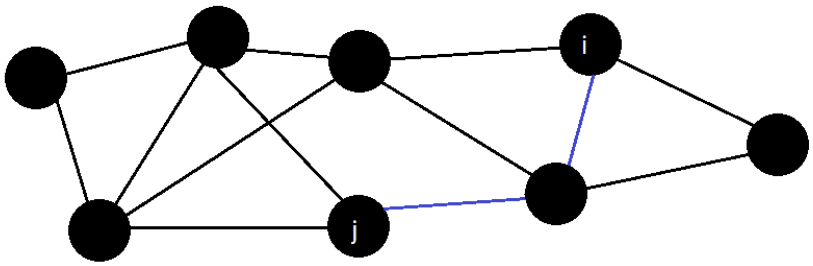


Figure 5: The path length from node i to node j, in this undirected network A, is two. The shortest distance between the nodes is the path coloured in blue. The degree centrality for node i and j is three.

been mapped, the node has a distance of $d + 1$[8].

2.5 Centrality

Centrality has been used to describe how important a specific user is for spreading information in a social media network[8]. There are different types of centrality, and choice of type should depend on the network we want to study and what information we intend to gather.

Degree centrality for a node, is the number of neighbouring nodes that a node can directly affect[8]. For a directed network, we need to pay special attention to the relation represented by the edge. It would be easy to draw the false conclusion that the degree centrality for a node in a directed network is the number of nodes out from the edge. However, for a Twitter network, it is the number of edges going in to the node, which is the number of followers a user has, k_i^{in}.

Closeness centrality for a node can be obtained by calculating the average distance from the node to all other nodes in the network[8]. This might give a more representative description than the degree centrality, because two users with the same follower count will have the same impact on their networks. With the closeness centrality, consideration is taken to how the node is connected to the rest of the network. The downside is that we need more information about the network to calculate the closeness centrality for an node, whereas to calculate the degree centrality we only need to have the nodes ego-network. To calculate the closeness centrality for a node i in a network, calculate the mean distance l_i: $l_i = 1/n \sum_j d_{ij}$, where d_{ij} is the geodesic path from i to j, then the closeness centrality is the inverse of the mean distance:

$$C_i = n/\sum_j d_{ij} \tag{3}$$

Betweenness centrality denoted by x_i describes to what extent a node is positioned between other nodes[8]. Nodes can become highly influential if a great number of geodesic paths include the node, as the loss of the node could increase the mean distance, these nodes can also control what information to relay through to the rest of the network. To calculate the betweenness centrality we first define n_{st}^i to be 1 if node i is on the geodesic path between nodes s and t, otherwise it will be 0[8]. the betweenness centrality can then be calculated using:

$$x_i = \sum_{st} n_{st}^i \tag{4}$$

This way of calculating betweenness centrality works fine, as long as there only is one geodesic path between s and t. In social network there is the possibility that more then one geodesic path exist between two nodes, and calculating the betweenness centrality for a node on the geodesic paths between s and t is not as straight forward as mentioned above. In this case, another method of calculating the betweenness centrality is required. To get around this problem, one simply takes the number of geodesic paths between two nodes and give

them a weighted value. First g_{st} is defined to be the number of geodesic paths between the nodes s and t, and n_{st}^i is defined to be the number of geodesic paths between s and t that pass through i[8]. The betweenness centrality for node i can be calculated by:

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{5}$$

The question of centrality is not as much a question of which method to use, since we are working with an ego-network the only suitable method is the degree centrality. The question that arises is which number do we use, the number of followers given by Twitter, or the number of followers after the criteria of activity and unlocked account is applied? The answer is of course, the number Twitter provides. The reason for this is because neither a locked account nor an account without recent status updates, our indication for an active account, prohibits Blondinbella from influencing these users. The user with the blocked account will still receive status updates from the user that they follow, and a user whom does not post any statuses could still be consuming other users statuses. However, when we want to study the amount of statuses and followers for users, it makes sense to not include users with locked accounts and users that do not post statuses.

2.6 Power law distributions and preferential attachment

As described in the introduction, Albert and Barabási suggested, in the paper: Emergence of Scaling in Random Networks, that nodes in social networks does not form edges to new nodes in a manner that could be described by an equal probability model, like the Watts and Strogatz model[6]. The Watts and Strogatz model also, did not describe what would happen when new nodes were introduced to the network, instead it focus on existing edges changing destination. The Albert and Barabási model on the other hand incorporates an explanation regarding new nodes and their introduction to the network based on preferential attachment, where nodes with a high degree had a high degree to generate new connections when new nodes were introduced to the network. The probability that a node, j, in a network m_0 will create a connection with a new node, i, is dependent on its connectivity, k_j, and the probability, Π that a new node will form edges to an already existing node i is[6]:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_i} \tag{6}$$

Every new node is allowed to create m edges, were $m \leq m_0$. After t steps, in which every step introduce a new node to the network, the network will consist of $t + m_0$ nodes and mt edges[6]. This network is going to have power law characteristics, and the probability that a node in the network has degree k follow a power law with $\gamma = 2.9 \pm 0.1$.

The rich get richer phenomenon arise from the fact that a node with high connectivity will have an easier time to create new edges when a new node is introduced to the network. This promotes nodes with an already high degree to increase their degree with time, while nodes with a low connectivity will have a low probability to create connections when new nodes are added to the network. Since every new node will increase the connectivity of nodes with an already high connectivity with a higher probability then nodes with a low connectivity[6], this will create a class of nodes with an extreme in-degrees. In social media, this would probably be best illustrated by pop stars and athletes, gathering millions of followers on Facebook, Twitter and Instagram.

The general degree distribution for a scale free network with preferential attachment follow the a power law function, $k^{-\gamma}$[6].
Newman suggest a way to calculate the power law coefficient[7]. Consider the distribution created by: $P(x) = C + x^{-\alpha}$ This function will produce a power law distribution as described above. Newman suggested that α be calculated through:

$$\alpha = 1 + n[\sum_i ln\frac{x_i}{x_{min}}]^{-1} \tag{7}$$

2.7 Assortive mixing in networks

The connections between nodes in a network might appear random and rather chaotic, especially for large networks. There might, however, be some reason to the way edges are connected the way that they are. There might be some characteristic that govern if a node will create a connection to another or not. If there exist such a characteristic, we say that the network has assortive mixing with regards to that characteristic.
In the paper Birds of a feather: Homophily in social networks written by McPherson et al. describe the prevalence of assortive mixing in social networks, and according to this paper, homophily is present in most kinds of social networks[9]. There is some evidence to support the statement that the more relations there are between two people, the stronger is the homophily, that is, the more relations there are between two people, the more alike they are[9]. McPherson describe two different kinds of homophily, as described by Lazarsfeld and Merton[10], the fist kind of homophily is status homophily which is based on formal, informal or ascribed statuses [9]. The second type of homophily is value homophily which is based on values, attitudes and beliefs[9]. On a social media platform it would be possible to measure assortive mixing based on characteristics of either status- or value homophily.
To measure assortive mixing, Newman proposed that a assortivity coefficient r be used[12]. The equation would differ depending on the characteristics studied, and Newman gave equations for calculating discrete(equation 8), scalar (equation 9) and a special case for when the characteristics studied are the degree of nodes (equation 10) [12]:
Discrete characteristics:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{8}$$

Here, e is an mixing matrix consisting of the fraction of edges that start or end in a node of type i [12]. for a undirected network : $a_i = b_i$. The following sum rule is satisfied by e:

$$1 = \sum_{ij} e_{ij}, \quad a_i = \sum_j e_{ij}, \quad b_j = \sum_i e_{ij} \tag{9}$$

Scalar characteristics:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_{xb_y})}{\sigma_a \sigma_b} \tag{10}$$

Here, as it was with the case of discrete characteristics, e is a mixing matrix consisting of the fraction of edges that connect edges of value x to edges of value y. σ_x and σ_y are the standard deviations for the distributions a_x and b_y [12]. Again, e satisfy the following sum rules:

$$1 = \sum_{xy} e_{xy}, \quad a_x = \sum_{y} e_{xy}, \quad b_y = \sum_{x} e_{xy} \tag{11}$$

Degree as an characteristic:
As stated previously, degree is a special case of mixing with regards to scalar characteristics. For a directed network the assortivity coefficient is calculated by:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j^{in} q_k^{out})}{\sigma_{in}\sigma_{out}} \tag{12}$$

Here, e_{jk} is the probability that a random edge leads to a node with in-degree j from a node with out-degree k, σ_{in} and σ_{out} are the standard deviations for the degree distributions[12].

2.8 Twitter, tweets and retweets

We define a tweet as a piece of information published to a specific user's Twitter feed through the tweet function on the Twitter platform. We define a retweet as the sharing of another user's status update through the retweet function on the Twitter platform. An example of a tweet can be found in figure 5.

Statuses are defined as the combined number of tweets and retweets a user has created within a given timespan. The tweet function allows users to create and contribute information to their network, while the retweet function allows for the user to spread information that already was on the platform, to their Twitter network.

A user can choose to follow another user by subscribing to their feed. Users can subscribe to other users as long as the user that gets followed has not locked their account. This subscription is not necessarily mutual, and because of this, we treat a subscription as a directed edge in networks concerning subscriptions on twitter.

Figure 6: This is a status update, also known as a tweet, created by the user qipedia. The tweet has 484 retweets. The retweet indicator is circled in red.

2.8.1 The tweet/retweet ratio

To tweet and to retweet are two fundamentally different ways of spreading information in a user's network. The tweet is new information, created by the user, and so the user is contributing to the total amount of information on the platform, and more directly to their network. For the follower of a user that tweets a lot, a lot of new information will be available from the tweets. For a user that only follow users that only tweet, and never retweet, the user would receive a lot of original information.

A retweet, on the other hand, is not original information. The user that retweets another users tweet has not created new information. With that said, the user might have introduced information that was new to their network. A user that only follow other users that only retweet, that user would receive a lot of information from users that is not part of his or her own ego network.

To test if a user is more prone to tweet or retweet, we construct a fraction; the tweet/retweet ratio, see equation 8.

$$\frac{Tweet}{Retweet + 1} \tag{13}$$

If the ratio is greater than one, the user tweets more than he or she retweets. If the ratio is lower than one, the user retweets more than he or she tweets. The plus one in the quota is there to avoid division by zero since a user might never have retweeted.

2.9 Nearest neighbours average connectivity

As discussed earlier, when measuring populations with regards to traits, where there is a majority and minority presence, correlation test results might be misleading if no consideration is taken regarding the underlying statistical distribution these traits adhere to. There might not be any correlation other than the one created by the statistical distribution. In this case, if we assume k_{in} adhear to a power law distribution, consideration must be taken in regards to how to account for the power law distribution when doing correlation tests[13]. For the case of k_{in}, where the test conducted is if users with a high k_{in} form connections with other users with a high k_{in}, There might be an influence from the power law distribution k_{in} adhears to. If we fail to recognise this, results might show a correlation were none actually exists. One such result would show that users with high k_{in} to a great extent follow users with lower k_{in} than themselves.
To examine if there is a non-trivial correlation with regards to connectivity, the nearest neighbours average connectivity,NNAC, for the network, $\langle k_{nn} \rangle$ is tested.

To calculate $\langle k_{nn} \rangle$, one can use equation 9: [13]

$$\langle k_{nn} \rangle = \sum_{k'} P_c(k' \mid k) \tag{14}$$

Were $P_c(k' \mid k)$ is the conditioned probability that a node with connectivity k is connected to a node with connectivity k'[13]. If this is independent of k, then there is no correlation between the nodes connectivity, if on the other hand, there is a dependency to k, then there is a correlation with regards to connectivity in the network[13].

3 Method

3.1 Users on Twitter and their representation in the adjacency matrix

For a Twitter user i, the subscriptions user i made on the platform, and the other users j, that subscribe to user i, their relation can be described with a adjacency matrix. Since the relation of subscriptions is directed,the adjacency matrix might be asymmetrical. The elements of the adjacency matrix will be:

$$A_{ij} = \begin{cases} 1, & \text{if j is subscibed to i:s feed} \\ 0, & \text{else} \end{cases} \tag{15}$$

To calculate k_i^{in}, Twitter provides $k_i^{in} = \sum_j^n A_{ij}$ for the huge A matrix consisting of the entire network. The same goes for the number of statuses for i.

3.2 Data collection

Data was collected with a program written in Python. The program utilised the Twitter API (Application Program Interface) to access the Twitter website and get information about accounts. Several purpose-build programs were created to collect specific data about the accounts. The data collected was; number of tweets, number of retweets and the number of followers a user had. To access the Twitter API, an application was created on the Twitter developer page. There is, however, a limit to the number of requests that a single application can make per 15 minutes. Because of this, data collection takes longer than one might expect. User tweet and retweet numbers are based on the users activity during the period from the 17th of February 2018 to the 19th of March 2018. A total of 13228 users' data was collected.

If all of Twitter could be gathered in one huge adjacency matrix, $A^{Twitter}$ the collection of users would follow this model:

For $\sum_j^n A_{Blondinbella,j}$, if $A_{Blondinbella,j}$ for node $j = 1$, collect user. Were j is an user on Twitter.

To test for assortive mixing, a sample of the users in the ego-network was randomly drawn. That is, if a user follows Blondinbella, and does not have an inactive or locked account, the user could be chosen. 486 users were tested between 8th of April and the 18th of April 2018. The chosen user id was tested against all other members of Blondinbellas ego network.

To be able to test if the chosen user was followed by anyone in the network, a registry of all users in the network with an active non-locked account was created. In this registry every user got assigned a text file. This file was filled with the id's of all the other users that specific user followed. The chosen user id, that had been randomly selected as described above, was tested against all these files, and if the user id was found, the user was followed by the user corresponding to the list. When this happened, follower count, tweet and retweet numbers was collected for the follower and the followed.

Using $A^{Twitter}$, this selection process can be described as:
For a user j, if user j is in $\sum_j^n A_{Blondinbella,j}$, include j to the set Followers of Blondinbella, now a considerably smaller adjacency matrix $A^{Blondinbella}$ can be constructed using the set Followers of Blondinbella, were:

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge from j to i} \\ 0, & \text{else} \end{cases} \tag{16}$$

This adjacency matrix might contain users f with $\sum_j A_{fj}^{Blondinbella} = 0$ and $\sum_i A_{if}^{Blondinbella} = 0$, that is, they do not follow anyone or get followed by anyone in the network. This is acceptable, and to be expected

The program would select a random row j in this matrix. If it finds a one in the A_{ij}:th position, it registers a relation from the user j to user i. The scalar quantity of the metrics studied would be registered.

3.3 Correlation test

To explore the possibility of predicting tweet or retweet behaviour, based on k_i^{in}, two different kinds of correlation tests were used. The Pearson correlation test and the Spearman rank correlation test.

The Spearman rank correlation test uses rank instead of value to perform the correlation calculation[15]. Because of this, outliers do not affect the result in the same way as in a Pearson correlation test.

The Pearson correlation test is the covariance for two sets of paired data divided by the product of the standard deviation of respective data set[8]. When testing for behaviour regarding tweets and retweets against degree, the test is conducted so that a user i with in-degree k_i^{in}, tweet count T_i and retweet count

RT_i, the tweet/retweet, $(T/RT)_i$ ratio is calculated the same way, are tested against each other, for the nodes in the ego network this is calculated as[14]:

$$r = \frac{cov(k, T)}{\sigma_{k^{in}} \sigma_T} \qquad r = \frac{cov(T, RT)}{\sigma_T \sigma_{RT}} \qquad r = \frac{cov(k, RT)}{\sigma_{k^{in}} \sigma_{RT}} \tag{17}$$

Due to the fact that the test compares the in degree and tweet/retweet behaviour, which is all discrete variables, for the same node in the network, the expected result would be that users with a low in-degree also show low values on T_i and RT_i.

When tests were conducted, a user j in the set consisting of Blondinbellas followers, was tested against themselves in the metrics described above.

3.4 Assortivity test

Two users that had the relation of follower and followed, if one user follow the other user, and both users were in $A^{Blondinbella}$, were tested. If a following relation was found, the number of followers for the user following was tested against the number of followers for the user being followed. To test for tweet, retweet, statuses and the tweet/retweet ratio, the number of tweets, retweets, statuses and tweet/retweet ratio for the follower was tested against the number for the user being followed.

Newman propose a way for calculating the assortivity coefficient r, to test for assortive mixing by scalar properties in the paper Mixing patterns in networks[12]. First, the relation between two nodes in i directed network is an edge from node j to node i, were k_j^{in} is x and k_i^{in} is y. From this a matrix e_{xy} is constructed, which is consisting of the fraction of edges that link an edge of value x to another with value y. In a perfectly mixed network, $y = x$ for all measures. The following sum rules apply:

$$1 = \sum_{xy} e_{xy} \qquad a_x = \sum_{y} e_{xy} \qquad b_y = \sum_{x} e_{xy} \tag{18}$$

Were a_x is the fraction of edges that start at value x, and b_y is the fraction of edges that end at value y. From the distributions of a_x and b_y, their standard deviations can be calculated as σ_x σ_y The assortivity coefficient r can then be calculated by equation 14:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_x \sigma_y} \tag{19}$$

Equation 10 can be used for all scalar properties, such as Tweet count, retweet count and tweet/retweet-quota, without having to modify the equation.

3.5 Nearest neighbours average connectivity

The nearest neighbours average connectivity is obtained by using $A^{blondinbella}$. For a user in $A^{Blondinbella}$, if $A_{ij} = 1$, user j follow at least user i and Blondinbella. Now, using $A^{Twitter}$, create the set: Users that j follow.

If user j is in both $A^{Blondinbella}$, $A^{Twitter}$ and $\sum_i^n A_{ij}^{Blondinbella} \neq 0$
Now, for every user j, were:

$$A_{ij}^{Blondinbella} = \begin{cases} 1, & \text{if j is subscibed to i} \\ 0, & \text{else} \end{cases} \tag{20}$$

For every 1 found in position $A_{ij}^{Blondinbella}$, add the user i to the set Users that j follow. For every user i in this set, gather the in-degree k_i^{in} supplied by Twitter and the size of the set: Users that j follow. Now the nearest neighbours average connectivity for the user j is calculated by:

$$NNAC_j = \frac{\sum_i^{Usersthatjfollow} k_i^{in}}{|\text{Users that j follow}|} \tag{21}$$

The $NNAC_i$ is compared to the in-degree for user j supplied by Twitter in a Pearson- and Spearman correlation test.

3.6 Data processing

The data collected was processed in Matlab. Both Spearman and Pearson correlation test were performed due to the prevalence of outliers. To test for tweet, retweet and the tweet/retweet quota against the user follower count, correlation tests were used.

4 Limitations

4.1 User activity and locked accounts

User data collection was restricted by a program designed to sort out specific accounts. First, inactive accounts were excluded. Here we define an account to be inactive if no new statuses had been created in 60 days. Users with zero statuses were also excluded. Another limitation the program was designed to account for is locked accounts. A locked account locks a user's feed, so that other users cannot see their statuses, unless the user approves that another user gets access to their feed. Since we will not be able to reach out to everyone with a locked account, instead they are removed.

From a total of 81026 followers, 67798 users were deemed not suitable by the criteria above.

4.2 Bots and bought followers

Bots are computer generated user accounts. These accounts can be used for several different purposes. They spread links to malicious websites and post disinformation[16] aimed to sway public opinion and increase a users follower count. The subject of bots in social media, their interactions and the result of their presence cannot be summarized in this rapport. However, we need to be aware of the fact that some users might be computer generated, and that their behaviour can affect how representative a study such as this one can be. We are after all trying to use the number of followers to predict tweet and retweet behaviour, these parameters can be manipulated by a bot, and that affects the validity of the study. However, care can be taken to present and analyse the data in such a way that the influence of bots is limited.

In a review article by Emilio Ferrara (et al.) describes typical behaviour of social bots[17]. They found that social bots tend to retweet substantially more than a normal user, 4 z-points for the bots compared to 1 z-point for normal users[17]. Respectively bots tend to tweet a lot less than normal users, -3 z-points for bots and around -0.5 z-points for normal users[17]. With this information correlations will be calculated for statuses, tweets, retweets and followers respectively. For correlations with retweets and statuses, since retweets are a part of statuses, we need to be aware of the potential presence of bots. For correlations with tweets, the result should be more reliable.

A user might not be aware of the fact that they are followed by bots, however, in some cases the user is very much aware of the fact that some of his or her followers are not all normal users. This is the case when a user purchases follower. There are markets where a user can purchase followers [16].

5 Results

The correlation and assortivity result for all characteristics can be found in appendix A.

5.1 Basic network information

The α value described in equation 7 was 1.2025. α is based on the in degree of Blondinbella. Figure 4 show the in-degrees for users j that subscribe to users i, on a log-log histogram. It is clear that middle of the histogram follow the description for law behaviour, with the tails deviating. As described by Newman, it is not necessary for the graph to follow the expected behaviour for the entirety of the range[8]. According to figure 5, there is a clear power law characteristics between 10^2 and 10^5, with small deviations at the end of the graph. The majority of the deviation can be found for lower values of k, something Newman claims to be typical[8].

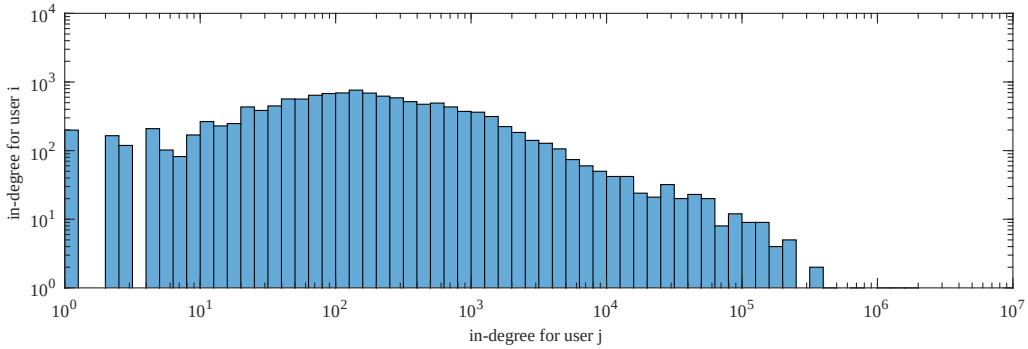


Figure 7: Histogram for the in-degree for a user j that follow a user i. Both axles are logarithms (base10).

Table 1 contain some basic statistics for users j following Blondinbella, sampled during this study.

	Followers	Tweets	Retweets	Statuses
Mean	1800.2	45.072	28.2955	73.3675
Median	136	4	2	9
Standard deviation	23916	167.091	149.9167	252.53

Table 1: Basic statistics for all users j that follow Blondinbella.

5.2 Tweet and retweet behaviour depending on follower count

No correlation was observed for the tweet/retweet ratio in either Person or Spearman correlation. A correlation was found between the follower count and the number of statuses, tweets and retweets, using the Spearman rank correlation. The *rho*-values were 0.4449, 0.442 and 0.3268. All correlations had a p-value less than 0.05. Correlations were also found between the number of tweets and retweets and the number of statuses a user had made, this is of course to be expected, after all, tweets and retweets make the statuses. Table

Characteristic	r	p
Tweets	.0817	< 0.05
Retweets	.0522	< 0.05
Statuses	.0851	< 0.05
$\frac{tweet}{Retweet}$	.0107	0.2165

Table 2: Pearson correlation for characteristics against in-degree

Characteristic	r	p
Tweets	.4622	0
Retweets	.3542	0
Statuses	.4656	0
$\frac{tweet}{Retweet}$	.0576	< 0.05

Table 3: Pearson correlation for characteristics against in-degree, all values are natural logarithm

Characteristic	rho	p
Tweet	.442	0
Retweet	.3268	0
Statuses	.4449	0
$\frac{tweet}{Retweet}$	-.0225	< 0.05

Table 4: Spearman correlation for characteristics against in-degree

5.3 Assortivity

The result from the assortivity test did not yield any Pearson correlation values (r-value) greater than -.1 or .1. However, Spearman correlation test yielded -0.2085 for follower count. The corresponding p-value was $5.0314 * 10^{-240}$, so the correlation is statistically significant. There was no apparent evidence to support assortive mixing, and therefore assortivity with regards to the tweet/retweet ratio. As Figure 5 shows the bivariate distribution for the number of followers for a user that is following another user, the x-axis, against the number of followers for the user that is being followed, the y-axis. Both datasets are logarithms (base10).

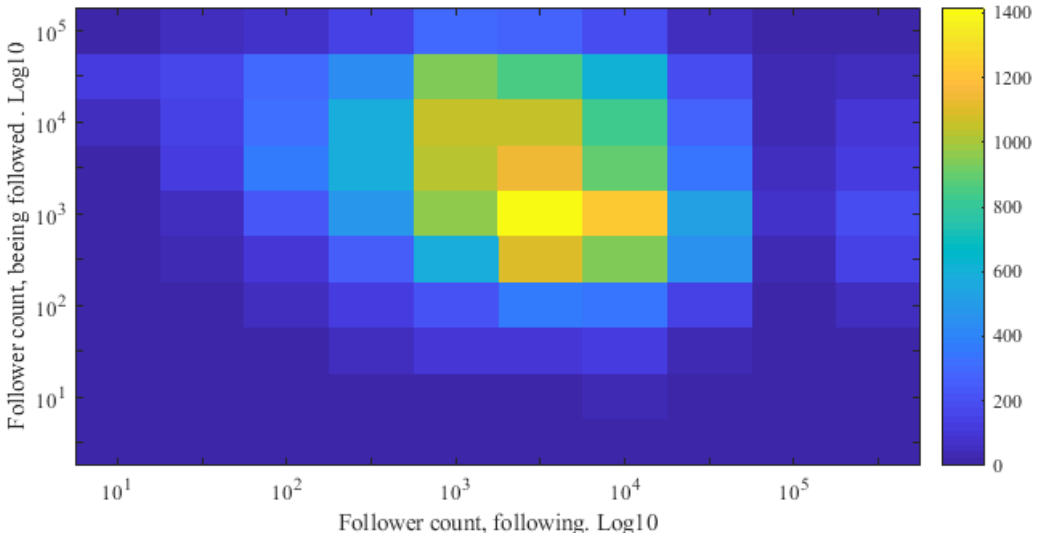


Figure 8: Bivariate histogram for the number of followers for a user that is following another user, against the number of users for the user being followed. Both datasets are logarithms (base10).

5.4 Nearest neighbours average connectivity

The results for nearest neighbours average connectivity can be found in Table 5.

Test	rho/r	p
Pearson	-0.0157	0.0847
Pearson log_{10}	-0.0735	< 0.05
Spearman	-0.0823	< 0.05

Table 5: Results for the NNAC-test,

6 Discussion

The tweet/retweet ratio does not correlate with the number of followers that a user has. A possible explanation for this could be the fact that a user might not have read through all of another users statuses before they decide whether to follow them or not. Perhaps other parameters are more important when users

decide who to follow. This also shows that we can not use the follower count and the tweet/retweet ratio to predict how a user will create and spread information to their network.

There is however a correlation between the number of followers a user has and the number of statuses that they have posted. We do not however know which one depends on the other, at this point the only conclusion that can be made is that users with a lot of followers tend to also have a lot of statuses. Further research could be conducted to find out if a higher follower count encourages a user to create more statuses, or if a user that keeps creating statuses eventually will accumulate a lot of followers by simply being active on the platform. Further studies should be conducted, in which users follower count and status count is monitored during several months.

There appear to exist a disassortivity regarding in-degree in $A^{Blondinbella}$. This means that users tend to follow other users with a lower in-degree than themselves. This can seem counter-intuitive; however, we must remember that this only is a sample of an network, and that the results might have differed if the entire Twitter network were to be tested. For this sample it might be tempting to say that users tend to follow other users with fewer followers that themselves. However, this might be due to the fact that we are testing in-degree, which have been shown to follow a power law degree distribution. As mentioned in the introduction, care must be taken when studying majority and minority groups. In this case we have a minority of the users that have a majority of the followers. The vast majority of users do not have that many followers, as can be deducted from the fact that the median follower count was 136 (Table 1). One other factor to take in to account is the nearest neighbours average connectivity. Since it did not give any indication of correlation with regards to connectivity, my stance is that the dissassortivity should be contributed to the underlying power law distribution and not user consciousness regarding the number of followers another user have. This might seem to disagree with previous research, however, this study was of an extremely small part of a vast network, and it should be argued that this result does not contradict previous research.

There is the case of accounts with lots of followers that does not produce any statuses. These accounts will, according to the degree centrality, still impact their ego-networks. So, care must be taken to make sure that the user in question create statuses of some kind, complete silence can not be considered to exert influence in social media.

An interesting observation was that a few users, that had a low follower count, ten followers or fewer, had a high number of tweets. This is however not reflected in the correlation coefficients since these users are a minority in the network. Of course, some of these users might be bots, however, some of the users that show this behaviour are not bots, this has been determent by visual inspection. This raises the question: why do you keep creating statuses? This is

also a possible field for continued research. Right now, it just appears as these users are screaming in to the void of the internet.

References

[1] Facebook Reports First Quarter 2018 Results. (2018). *Retrieved from https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-First-Quarter-2018-Results/default.aspx*, Last read 2018-05-09

[2] Six Degrees of Separation. (2018). Retrieved from http://www.oxfordmathcenter.com/drupal7/node/655 Last read 2018-05-10

[3] Travers J, Milgram S. *An experimental study of the small world problem*, *Sociometry* 1969;32:425-443.

[4] Watts D, Strogatz S. *Collective dynamics of 'small-world' networks*, *Nature* 1998;393:440-442.

[5] Bonacich P, Lu P. Introduction to mathematical sociology. Princeton: Princeton University Press; 2012

[6] Barabási, A. Albert, R (1999). Emergence of Scaling in Random Networks. Science, 286(5439), 509-512. doi: 10.1126/science.286.5439.509

[7] Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), pp.323-351.

[8] Newman, M. (2010). Networks. Oxford University Press.

[9] McPherson, M., Smith-Lovin, L., Cook, J. (2001). Birds of a Feather: Homophily in Social Networks. Annual Review Of Sociology, 27(1), 415-444. doi: 10.1146/annurev.soc.27.1.415

[10] Lazarsfeld PF, Merton RK. 1954. Friendship as a social process: a substantive and methodological analysis. In Freedom and Control in Modern Society, ed. M Berger, pp. 18–66. New York: Van Nostrand

[11] Blau, P., Blum, T. and Schwartz, J. (1982). Heterogeneity and Intermarriage. American Sociological Review, 47(1), p.45.

[12] M. E. J. Newman. *Mixing patterns in networks*, *Physical Review* 2003;67.

[13] Pastor-Satorras, R., Vázquez, A., Vespignani, A. (2001). Dynamical and Correlation Properties of the Internet. Physical Review Letters, 87(25). doi: 10.1103/physrevlett.87.258701

[14] Basic Concepts of Correlation — Real Statistics Using Excel. (2018). Retrieved from http://www.real-statistics.com/correlation/basic-concepts-correlation/ as read 2018-10-30

[15] Alm, S. Britton, T. (2008). Stokastik. Stockholm: Liber.

[16] Stringhini, G., Egele, M., Kruegel, C., Vigna, G. (2012). Poultry markets. ACM SIGCOMM Computer Communication Review, 42(4), 527. doi: 10.1145/2377677.2377781

[17] Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A. (2016). The rise of social bots. Communications Of The ACM, 59(7), 96-104. doi: 10.1145/2818717

A Appendix

A.1 Corretaltion test results

	Followers	Tweets	Retweets	Statuses	Tweet/retweet
Followers	1	0.0817	0.0522	0.0851	0.0107
Tweets		1	0.2670	0.8202	0.1282
Retweets			1	0.7703	-0.1483
Statuses				1	-0.0032
Tweet/retweet					1

Table 6: Pearson correlation values, r-values

	Followers	Tweets	Retweets	Statuses	Tweet/retweet
Followers	0	$4.9580 \cdot 10^{-21}$	$1.8562 \cdot 10^{-9}$	$1.1360 \cdot 10^{-22}$	0.2165
Tweets		0	$9.8468 \cdot 10^{-215}$	0	$1.2875 \cdot 10^{-49}$
Retweets			0	0	$5.7801 \cdot 10^{-66}$
Statuses				0	0.7115
Tweet/retweet					0

Table 7: Pearson correlation values, p-values

	Followers	Tweets	Retweets	Statuses	Tweet/retweet
Followers	1	0.4622	0.3542	0.4656	0.0576
Tweets		1	0.4519	0.8591	0.4393
Retweets			1	0.766	-0.4407
Statuses				1	0.0311
Tweet/retweet					1

Table 8: Pearson correlation values, r-values, natural logarithm

	Followers	Tweets	Retweets	Statuses	Tweet/retweet
Followers	0	0	0	0	$3.4499 \cdot 10^{-11}$
Tweets		0	0	0	0
Retweets			0	0	0
Statuses				0	$3.4506 \cdot 10^{-4}$
Tweet/retweet					0

Table 9: Pearson correlation values, p-values, natural logarithm

	Followers	Tweets	Retweets	Statuses	Tweet/retweet
Followers	1	0.442	0.3268	0.4449	-0.0225
Tweets		1	0.3318	0.8192	0.3407
Retweets			1	0.7084	0.6769
Statuses				1	-0.1202
Tweet/retweet					1

Table 10: Sperman correlation values, rho-values

	Followers	Tweets	Retweets	Statuses	Tweet/retweet
Followers	0	0	0	0	0.0096
Tweets		0	0	0	0
Retweets			0	0	0
Statuses				0	$8.6094 \cdot 10^{-44}$
Tweet/retweet					0

Table 11: Spearman correlation values, p-values

A.2 Assortivity test results

Characteristics	r	p
in-degree	-0.0474	< 0.05
Tweet	0.0234	< 0.05
Retweet	0.0108	0.0897
Statuses	0.0291	< 0.05
Tweet/Retweet	-0.0057	0.3743

Table 12: Pearson correlation values

Characteristics	rho	p
in-degree	-0.2085	< 0.05
Tweet	0.02760	< 0.05
Retweet	-0.0051	0.4206
Statuses	0.0826	< 0.05
Tweet/Retweet	0.0122	0.0558

Table 13: Spearman correlation values