Analysis of the early evolution of the Twitter (X) social network in Sub-Saharan Africa

Watson Levens (watsonlevens@udsm.ac.tz)
University of Dar es Salaam
David J. T Sumpter
Uppsala University
Egbert Mujuni
University of Dar es Salaam
Idrissa Said Amour
University of Dar es Salaam

Research Article

Keywords: Twitter networks, power-laws, degree distributions, clustering coefficients

Posted Date: December 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3713138/v1

Additional Declarations: No competing interests reported.

Analysis of the early evolution of the Twitter (X) social network in Sub-Saharan Africa

Watson Levens[1], Egbert Mujuni[1], Idrissa Said Amour[1], and David J. T. Sumpter[2]

[1] *Department of Mathematics, University of Dar es Salaam, Tanzania*
[2] *Department of Information Technology, Uppsala University, Sweden*

In sub-Sahara African countries the proportion of the population using Twitter (now known as X) and other social media networks is growing. Understanding these networks allows us to understand changes in how people interact with each other, share information and carry out economic transactions. We hypothesise that the networks of sub-Sahara African Twitter networks have statistical properties consistent with being built on a principle of self-organisation: where decisions about who to connect to are less influenced by large commercial actors (as they might be in Europe, USA and other parts of the world) and are instead on the basis of local interactions between individuals. To test this hypothesis, we first collected data on the Twitter network of users in Tanzania. We found that the degree distribution followed a power-law with degree close to 2. We calculated path lengths, clustering, and assortativity of mixing for this network, as well as identifying the most influential users using eigenvector centrality. We then tested the degree to which these measurements were consistent with a variation of a friend-of-a-friend model of network attachments: where links are formed by individuals who join a network first identifying one individual at random (a friend) and attach to them and then choose n_q people who the initial individual follows and attach to each of them with probability q. We found that for $q = 1$ and $n_q = 40$ the model reproduces many aspects of the Tanzanian network, including the degree and clustering distribution. This model is not consistent with, for example, the USA or Japanese Twitter networks. Taken together, the model and its comparison to data from different real world network, supports the self-organisation hypothesis: a rule under which new members of the network connect to a random person and 40 people they follow reproduces many aspects of how the Tanzanian Twitter network has grown.

Keywords: Twitter networks, power-laws, degree distributions, clustering coefficients

Introduction

Social networks such as Twitter (throughout, we will refer to this network, now called X, with the name it had when this study was conducted), Facebook, Instagram, WhatsApp, WeChat, and LinkedIn have become more accessible and central to all aspects of human life. In their very diverse functionalities, these networks offer powerful new ways for

their users to connect, share real-time information, and influence the world. This world's traditional news paradigm shift to online communication has attracted a growing body of literature interested in understanding social networks' structural properties and growth processes [1, 2, 3]. This work can be traced back to 1998, when Duncan J. Watts and Steven Strogatz [4] proposed a family of small-world networks, characterized by short average path lengths and high clustering. Many large scale sparse networks such as internet, neuronal networks, the power grid and friendship human networks are examples of small-world networks. Another important form of network, scale-free networks have highly heterogeneous degree distributions, characterised by a power-law [5, 6, 7, 8]. Such networks can be generated by a process of preferential attachment, in which new nodes connect preferentially to the more highly connected nodes. Scale-free networks have been observed in online social networks [7] and elsewhere [9, 10, 11, 12]. Care is required here, however, because strict scale-free networks are rare in nature and in particular social networks are at best weakly scale-free [13].

In searching for a simple model to explain the growth of social networks, Levens et al. [14] have recently developed the friend-of-a-friend model for generating networks. The mechanism underlying this model is based on the idea that *a friend of your friend is your friend*. The evolution process of the model allows a new individual joining an existing network to select a friend randomly and establish a friendship with probability p. In the second stage of the model, the new person chooses one among the new friend's friends and connects with an independent probability q. This mechanism forms a network with a power-law degree distribution and small-world network clustering coefficient. When p is very small and $q = 1$, the mechanism produces super-hub networks.

One question we look at here is whether a variation of the friend-of-a-friend model is consistent with various Twitter networks. The interactions in the model are self-organised, in the sense that decisions on who to follow are made entirely through local recommendations from others. To what degree do real-world networks also follow this pattern? Previous studies show that Twitter networks are typically both small-world and weakly scale-free. For example, researchers working for Twitter (and thus with access to the entire network) were able to characterize the Twitter networks for USA, Japan and Brazil [15]. They observed a weakly scale-free degree distribution of connections. They also found that the shortest path length and clustering coefficient were within the limits of those expected for a small-world network. The Twitter study found no significant differences between the degree distributions of Twitter networks for Brazil, Japan and the USA, but the clustering coefficient was different for Japan and USA, where there was a peak at around degree $k = 100$ for USA and $k = 1000$ for Japan. Brazil Twitter network had clustering coefficient monotonically decreasing with the degree. In another study, though, for Purdue University students in the USA, it was found that the degree distribution had a power-law exponent $\alpha \approx 2.29$ and the average clustering coefficient C_T was 0.15 [1] and had properties largely consistent with a small-world network.

One aspect that is lacking in all the studies of Twitter and other social networks in sub-Saharan Africa. Usage of the internet and smartphone is still growing in these countries,

although the rate of usage across the world remains unbalanced [16]. Although the Twitter platform has users distributed and connected worldwide, sub-Sahara Africa is underrepresented. India and sub-Saharan Africa lag behind while countries in the western world, have reached an almost saturated level. Tanzania is listed as having the least internet and smartphone users worldwide [16].

We might then expect Twitter networks of sub-Sahara Africa countries to more closely reflect the self-organised principles underlying many network growth models. In contrast to the USA, for example, these countries have fewer large multinational companies shaping the network, meaning that decisions on who to follow are more likely to be shaped by independent actions of individuals. As such, we hypothesise that the Twitter network in sub-Sahara Africa is more likely to be characterised by the friend-of-a-friend model. To test our hypothesis we focus on Tanzania, characterising the social network on Twitter and comparing it to other networks.

Methods

Data collection and cleansing

Our aim was to better understand structural properties of Tanzanian Twitter network. To this end we collected data in two phases. First we identified a list of prominent people in Tanzanian politics from the official website for the parliament [17]. Between September 15, 2018 and November 16, 2018 we reviewed and identified those individuals' Twitter profiles and made a list consisting of 120 Twitter users. A personal Twitter account was then used to follow all of these people. To access the Twitter data, we registered with the Twitter Application Programming Interface (API) and we collected data for the selected Twitter users. Of the 120 Twitter users, 74 had mutual connections to other users and were thus included in our study as active prominent people in Tanzania. In other words, isolated accounts were ignored and we studied only the largest connected component.

The second phase was to collect the network of the Twitter accounts (not necessarily people) of followers of prominent people identified in phase one. We constructed a network that included all the accounts which followed the prominent people. This was 1,310,615 accounts. We then downloaded a list of accounts followed by each of these accounts. Since boundaries of the Twitter network are hard to define, as almost every person is eventually connected to almost everyone else in the world, we limited our study to only followers of prominent people identified in phase one. In other words, if A is a follower of any prominent person in Tanzania, then B is considered followed by A if and only if B is also a follower of any prominent person in the same country. This assumption was imposed to enable us to find users who are likely to be actively involved in Tanzanian society and to limit the study to (mostly) within the boundaries of the country. When the download was completed, this gave us a total of 1,514,435 accounts.

Because of the limitations imposed by Twitter where every access token is limited to perform 15 calls every 15 minutes, it took four months (between November 22, 2018

and March 11, 2019) to collect the full dataset. To avoid spending too many months collecting data we decided that we would download at most 2000 followed individuals for each account. Only 51 of these followed more than 2000 people, so we believe that this restriction did not greatly influence our results. Python programming language, both the networkx library and our own code on sparse matrices, was used in processing and analysis of data.

Measuring structural network properties

We consider Twitter network as a directed graph $G = (V, E)$, that consists of a set of nodes called twitter users or Twitter accounts, V , and a set of edges, $E \subseteq V \times V$, called links, which connects pairs of accounts through follower-following relationships. In evaluation of network measures and metrics, an adjacency matrix, $\mathbf{A}$, for all users was created. For the n accounts this is an $n \times n$ matrix with elements a_{ij} such that

$$a_{ij} = \begin{cases} 1, & \text{if user } i \text{ is a follower of user } j, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Note that $a_{ij} = 1$ does not necessarily imply $a_{ji} = 1$, since individuals do not always mutually follow each other on Twitter. To quantify and get more insights of the network topology, we measured common useful structural characteristics, including degree distribution, clustering coefficients, the average path length and network assortativity. We now describe these in detail.

Degree distribution

Degree distribution, $P(k)$, is one of the most commonly used measures to characterize connection of nodes in a network. It is defined as the fraction of nodes in the network with k edges. For directed networks like Twitter there are distributions for in-degree and the out-degree. The in-degree distribution of the Twitter network is then defined as the probability that one randomly chosen Twitter account has k followers in the network. The out-degree is the number of accounts followed by a randomly chosen Twitter account. The degree distribution of many real networks [9, 10, 11, 12, 18] follows power-law distribution, such that the probability that an account follows k accounts is $P(k) = Ck^{-\alpha}$, where $k \in \{0, 1, 2, 3, ..., n-1\}$ for a network of n nodes, the constant C is determined by normalization and α is a scaling parameter. In this study, we plotted and fitted in-degree data using a power-law with exponential cut-off,

$$P(k) = k^{-\alpha} e^{-\beta k}, \tag{2}$$

where a cut-off degree β for our data is approximated to be 400000. Logarithmic binning was applied to scale our data.

An approximated value of α was determined by the Maximum Likelihood Estimate method (MLE) as described in [19],

$$\alpha \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{k_i}{k_{\text{min}} - \frac{1}{2}} \right]^{-1}. \tag{3}$$

We used this formula because it is considered to give good results in cases where $k_{\text{min}} \geq 6$ [19].

Clustering coefficient

Clustering coefficient is a density metric that quantifies the transitivity relation of nodes connections in the network. It is calculated as the ratio between the number of triangles in the network and the number of nodes connected by two edges (number of both open and closed triplets) [6]. A triangle in the case of network analysis includes all three closed triplets or loops of length three. Therefore the global clustering coefficient (C) is simply calculated as,

$$C = \frac{3 \times \text{Number of triangles}}{\text{Number of triplets (open and closed triplets)}}. \tag{4}$$

It is also possible to calculate local clustering coefficient C_i [4] which provides local information of how triangles are placed around different vertices of the network,

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{5}$$

where $\frac{k_i(k_i-1)}{2}$ is the maximum number of possible edges of vertex i whose degree is k_i and E_i is the number of all edges that exist among all first neighbours of vertex i or the number of triangles passing through vertex i. The global clustering coefficient C_{WS} resulting from Eq. 5 is different from that given by Eq. 4:

$$C_{WS} = \frac{1}{n} \sum_{i=1}^{n} C_i, \tag{6}$$

where n is the number of vertices in the network.

In Twitter social networks, clustering coefficient quantifies the situation that if person A is a follower of person B and person B is a follower of person C, then person A and C are more likely to be followers of each other. Although Twitter is a directed network, computations of local clustering coefficients were carried on undirected network using Eq. 5. We performed an approximation by assuming undirected network and only considering a random subset of the vertices at each iteration.

Average shortest path length

Average shortest path length (L) is a crucial quantity when categorising a network class, since it is this quantity which decides to whether a network belongs to small-world networks or not. The Average shortest path length L is defined as,

$$L = \frac{\sum\limits_{i,j \in V} d(i,j)}{n(n-1)}, \tag{7}$$

where $d(i,j)$ is the distance from node i to j, V is the set of nodes in the network, and n is the number of nodes in the network. The function in NetworkX for very large networks requires long execution times [20] and there are no efficient algorithms which can compute average shortest path of directed complex networks (since it requires finding shortest paths for every node).

Given this challenge, we used the following approach to compute an average shortest path of the network. Firstly, NetworkX was used to compute the shortest path length of the largest connected component of phase one data which is a network of 74 mutually connected Twitter accounts in Tanzania politics (phase 1 network). This average shortest distance was 1.64. In the second step, we noted that it takes an average of 2 steps for a user of phase 1 network to reach another user of phase 2 network (all other users, see methods for details). It is then clear that the average distance from an account of phase 1 network to account of phase 2 network is approximately equal to 2. This is possible with the assumption that the network is undirected and all accounts of the network are connected to each other.

Assortativity mixing

Assortative mixing is a tendency of nodes in a network to connect or associate with other nodes that are similar to them in some way. In social networks, assortative mixing typically relates to the tendency of people to form connections with others who share similar attributes, such as age, education, income, or other characteristics. In this work we use assortativity coefficient r to measure whether nodes with high degrees (hubs) preferentially connect to other high-degree nodes or if they connect more randomly. r is calculated from

$$r = \frac{\sum\limits_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=0}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=0}^{n}(y_i - \bar{y})^2}}, \tag{8}$$

where x_i and y_i are the connected nodes indexed with i, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. If $r > 0$, the network is assortative, meaning nodes tend to connect to nodes with similar degrees. If $r < 0$, the network is disassortative, indicating that high-degree nodes tend to

connect with low-degree nodes. If $r \approx 0$, the network has no strong degree-degree correlation, implying a more random mixing of nodes. Assortative mixing can be examined in directed networks by considering degree types of nodes at either end of an edge involved in the assessment. The coefficient r for the degree pairs include: assortativity between in-degrees $r(\text{x-in}, \text{y-in})$, in-degree and out-degree $r(\text{x-in}, \text{y-out})$, out-degree and in-degree $r(\text{x-out}, \text{y-in})$ and assortativity between out-degrees $r(\text{x-out}, \text{y-out})$.

Eigenvector centralities

Eigenvector centrality is a quantity that measures the influence of a node in a network. It quantifies the situation where a node i on the network can be famous or influential if connected to many users, high-status users, or both on the web. In this work, we use networkX [20] to calculate the eigenvector centrality and determine the Tanzanian central Twitter accounts of the network. To be able to compare the groups of influencers such as politicians, business companies, etc, we align the eigenvector centralities by normalizing the eigenvectors of each group as follows: Let $x_i \in X$, where X is a set of eigenvector centralities of group of n influencers and x_i is an eigenvector centrality of a twitter account i. Then, a normalised eigenvector e_j is given by

$$e_j = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{9}$$

Model

To predict the observed statistics of the Tanzania Twitter network we created an extended version of the friend-of-a-friend model [14], described in the introduction, as follows. At each time step, a new person joining the network establishes a friendship with a randomly selected individual. For each person selected, a total of n_q neighbours are chosen randomly, and with a probability q, the new person links with each of the n_q neighbours. When $n_q = 1$ this is equivalent to the original friend-of-a-friend model with $p = 1$. The parameter n_q is introduce to allow us to tune the expected degree $E[k] = qn_q + 1$ of the model to any value. For the original model this degree is at most 2.

The network is initialised with $m_0 = n_q + 2$ nodes, where all the nodes are connected to each other, so that each node has degree $k = n_q + 1$. This ensures there are sufficient existing connections for the network for new nodes to connect to. Following methods outlined in [14], we can show that the total rate of increase of degree k is given by

$$\frac{dk}{dt} = \frac{1}{t} + qn_q \frac{k}{(1+q)t}. \tag{10}$$

The first term on the right here corresponds to the initial random attachment to a friend. The second term arises from the n_q friend of a friend attachments, which increase in proportion to k.

The probability density function for any time t_i falling within the time interval $[t_i, t]$ of which the new node joins the network is given by the uniform distribution

$$P(k,t) = \frac{1}{E[k]+t} = \frac{1}{1+qn_q+t}. \tag{11}$$

Using Eqs. 10 and 11 and again following the mean-field approach [14] for approximating the degree distribution $P(k)$, we obtain

$$P(k) = \left(1+\frac{1}{qn_q}\right)\left(2+qn_q+\frac{1}{qn_q}\right)^{(1+\frac{1}{qn_q})}\left(k+1+\frac{1}{qn_q}\right)^{-(2+\frac{1}{qn_q})}. \tag{12}$$

In the limits of very large k, the degree distribution exhibits the extended power-law given by

$$P(k) \sim \left(k+1+\frac{1}{qn_q}\right)^{-(1+\frac{1}{qn_q})} \sim k^{-(2+\frac{1}{qn_q})}, \quad \text{for all} \quad k. \tag{13}$$

From Eq. (13), the scaling parameter α can be computed using the exponent of k. Thus,

$$\alpha = 2+\frac{1}{qn_q}. \tag{14}$$

Eq. 14 shows that the power-law exponents of the model lies in the interval $[2, \infty]$. Note that this is an approximation, which works well [14, 21, 22, 23].

Contrary to the models that connect to both the friend and all friends' neighbours [21, 22, 23], the proposed model generates networks with mean in-degree $E[k] \geq 1$. The most interesting part of the proposed mechanism is that it is prevalent in many social networks, including Twitter and Facebook. On Twitter, for example, a person can randomly choose one person to follow and simultaneously be a follower of several others, followed by the selected person. Taking into account the mean degree of empirical social networks, which is $E[k] \geq 1$, and since our earlier proposed model [14] generates a directed network with $E[k] \leq 2$, we use the proposed model to predict the empirical Tanzanian Twitter network which has a mean in-degree close to 40. We also compute clustering using the undirected version of the model.

Results

Connected components of the network

The Tanzania Twitter network consists of 1,514,435 accounts, connected via 61,175,824 edges. Thus, on average, each Twitter account on the network follows 40.4 users. The network has 707,625 strongly connected components. 97% Twitter users in the network formed a weakly connected component. The largest strongly connected component of the network contains 806,811 Twitter accounts which is about 53.3% of all users. The largest

weakly or strongly connected component is expected to be a recognizable network of famous people, companies, music groups, sports team and newspapers, within which many of the most central members of the component follow each other.

Analysis of degree distribution

To get an insight of a nature of distributions exhibited on the Twitter network, we looked at variant variables across users which includes in-degrees and out-degrees. In order to look at how network accounts were distributed on the Twitter network, we started by plotting a frequency distribution of number of followers which are represented by incoming degree on a log-log plot shown by scattered blue points in Figure 1A. We see that the number of followers is low for most of the Twitter accounts, but there are a few users with very high number of followers. We then fitted our plot by probability distribution function of truncated power-law distributed variable. The minimum value in our analysis was set to $k_{\mathsf{min}} = 10$ and a value of scaling parameter α is 1.91.

The degree distribution of the out-going degrees is shown in Figure 1B. The most obvious difference between these distributions is that the in-degree follows power-law but the out-degree distribution does not. The setting in data collection of limiting data collection to a t most 2,000 followed Twitter accounts could partly account for an out-degree probability distribution of this kind. More importantly, however, it is also explained by the fact that very few users follows more than 2,000 Twitter accounts, while quite a few popular accounts (with more than 2,000 followers) do exist.

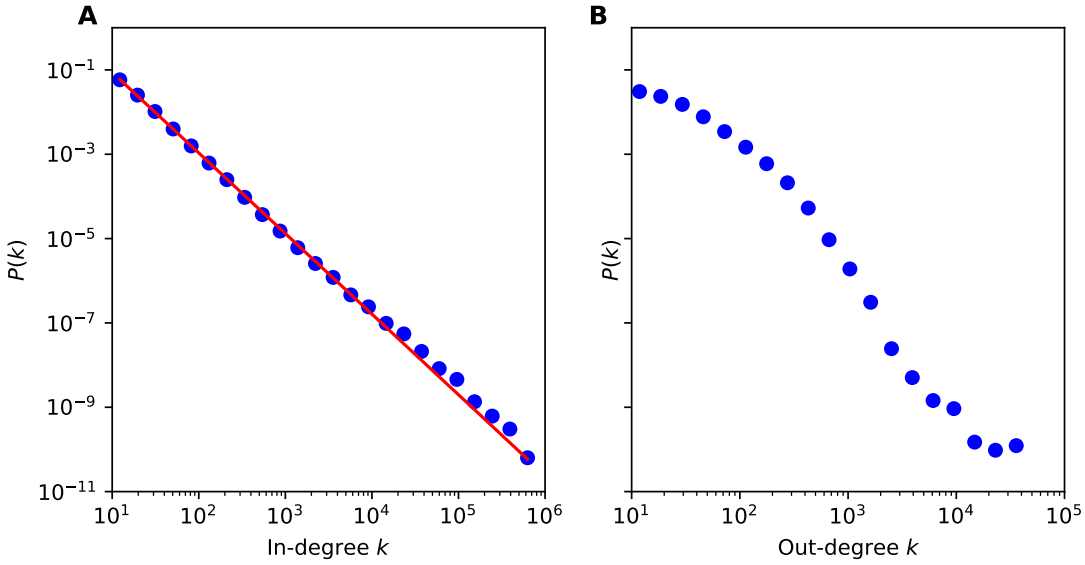

FIGURE 1: Degree distributions and power-law with exponential cutoff-fit. Panel A is degree distribution for Twitter followers (in-degree) data (blue scattered points) fitted with a normalized power-law with exponential cut-off distribution (red line). Panel B is an out-degree distribution for Tanzania Twitter data.

Clustering coefficient results

Clustering coefficient serves as a measure of how well the neighbours of twitter users are interlinked. The average clustering coefficient of the network computed using Eq. 5 is 0.2003, meaning that, of 1,514,435 Twitter users, only 303,341 (20%) follow each other. As shown by scattered blue points in Figure 3B, Twitter accounts with higher connections exhibit lower clustering coefficients, i.e. the clustering coefficient distribution is almost monotonically decreasing. The figure also shows that only 7 % of the users possess zero clustering coefficient and 93 % have greater than zero clustering coefficient. This means that most of the Twitter user's in Tanzania have at least two followers who in turn follow each other.

Average path length

The average of all-pairs shortest-path length of the network is the sum of single-source shortest-path length of each node in the mutual Twitter network and is estimated to be 3.64 (i.e. $1.64 + 2$ as described in Methodology). Based on number of Twitter users, United states is the leading country [24] and so it might be expected the Tanzania network

connectedness structure to be less mature than that of the US. However, based on the average shortest path length, Tanzania Twitter network is as tightly connected as that of US. This may be a consequence the way the Tanzania network data was collected. In data collection, we set a hub of most prominent people in Tanzania to be the source of obtaining other nodes linked to it and this imply that every node in the network is already connected with at least one other node. The Twitter network of people in Tanzania contains several connected components and every node has at least 1 degree. But in all cases, the shortest path lengths are still short enough to make the networks small world.

Assortativity mixing

The assortativity coefficient for nodes with both incoming connections is $r(\text{x-in}, \text{y-in})$ is 8.37×10^{-4}, suggesting a slight tendency for nodes with similar incoming degrees to connect. For nodes with incoming and outgoing connections $r(\text{x-in}, \text{y-out})$, the coefficient is $7.42 \times 10^{-}5$, indicating a weak tendency for nodes with high incoming connections to connect with other high in degree nodes and for nodes with low incoming connections to connect with other low in degree nodes. On the other hand, nodes with high outgoing connections tend to avoid connecting with nodes with high incoming connections $r(\text{x-out}, \text{y-in}) = -6.30 \times 10^{-}3$, and similarly for low degree. As shown by $r(\text{x-out}, \text{y-out}) = 4.82 \times 10^{-3}$, a minor bias exists for nodes displaying a similar level of outgoing degree to connect. Generally, the mixing pattern of Twitter users in Tanzania is very weak enough to be a random nature.

Central Twitter accounts in Tanzania

As might be expected, the Twitter network for Tanzanians consists of accounts for ordinary people, politicians, business companies, music artists, media outlets, journalists and more. Figure 2 shows the top 100 Tanzanian Twitter accounts with the highest eigenvector centrality of the network studied. Out of 100 most central accounts, 48 are politicians, 26 are musicians, 14 are journalists and 10 are news media and companies accounts. In this sample, there are only two companies with Twitter accounts which implies that in Tanzania, businesses had very low influence on Twitter at that time.

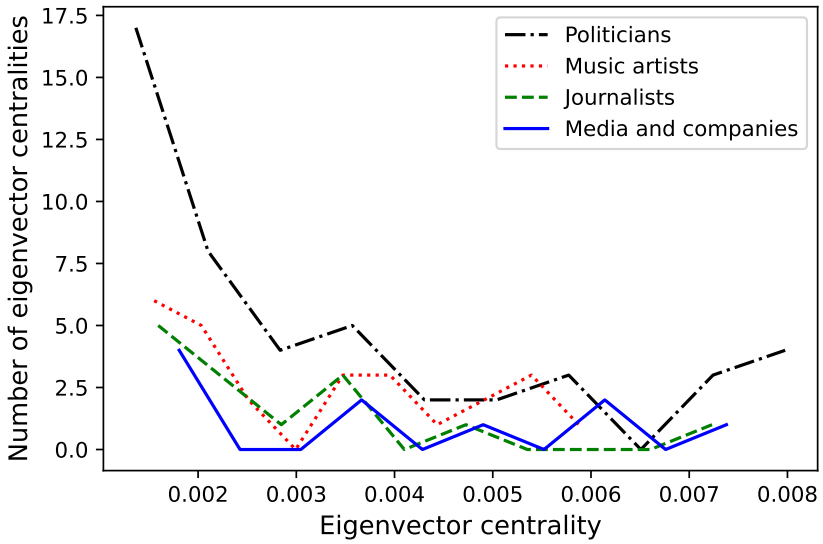

FIGURE 2: The counts of eigenvector centralities for 100 most central Twitter accounts in Tanzania. The eigenvectors for each group of users are normalized.

Comparison of the model with the Tanzania Twitter data

We performed simulations of the model described above and compared how the relevant global and local topological features of the networks relate to that of Tanzania Twitter network. We are interested in investigating how the different model parameters could produce a Twitter network. This has been done by studying the behaviours of the degree distribution $P(k)$, the average shortest path length L, and the network clustering coefficient C_T.

Figures 3A and B show the fitting of the in-degree distribution and clustering coefficient of the model to that of the Tanzanian Twitter, respectively. The data for the model is generated with $q = 1$ and $n_q = 40$. The degree distribution and clustering coefficient fit well the Twitter data. Additionally, the values of scaling parameter α and average clustering coefficient C_T shown in Table. 1 are nearly equal to that of Tanzanian Twitter network.

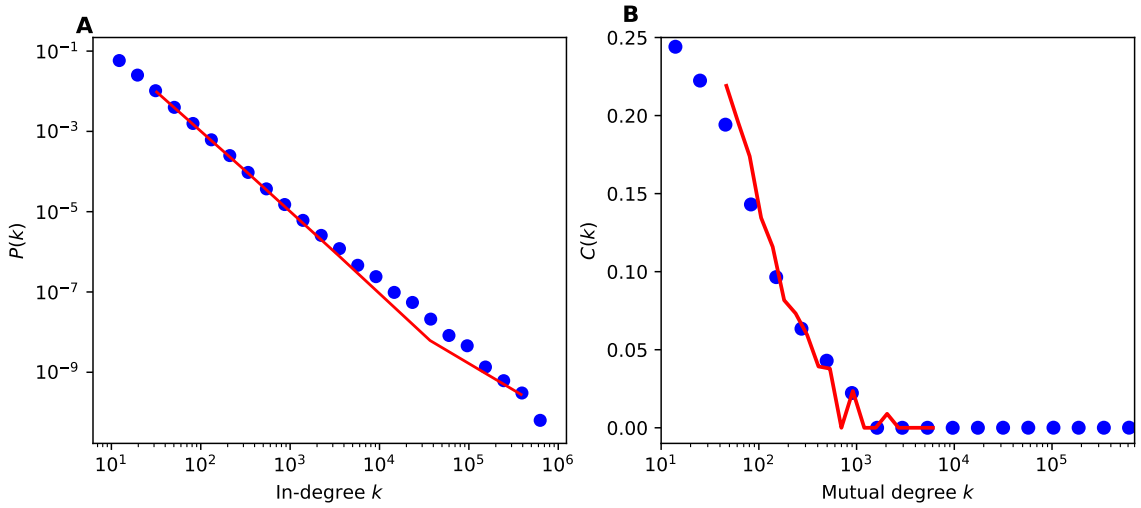

FIGURE 3: Fitting the degree distribution (panel A) and clustering coefficient (panel B) for the model to the Tanzania Twitter data. Scattered blue points are the empirical data and the red lines are the predictions. Model parameters used in each panel are $q = 1$, $n_q = 40$, $N = 1514435$ and mean in-degree $m = 41$.

Table 1: The basic properties of Tanzania Twitter network and the model.

Network element	Tanzanian Twitter	model
Network size (N)	1514435	1514435
Average clustering coefficient	0.19	0.18
Mean in-degree	40.4	41
Scaling parameter (α)	1.91	1.95
Power-law degree distribution	Yes	Yes

Table 2: Statistical properties of Twitter networks for Tanzania in 2018/2019 and Brazil, Japan, and the USA in 2012.

Network element	Tanzania	Brazil	Japan	USA
In-degree scaling parameter (α)	1.91	1.30	1.35	1.33
Average shortest path length (L)	3.64	3.78	3.89	4.37
Average clustering coefficient (C_T)	0.19	-	-	-

Discussion

There are some similarities and some differences between the Tanzania Twitter network and other social networks. The largest strongly connected component of Twitter network for people in Tanzania (53.30%) is slightly lower than that of Twitter follower network of all active users (68.70%) [15] . The in-degree distribution of the Tanzanian network does not deviate much from that of the entire Twitter network [15] and the network has small world properties [4]. Tanzania data shows a straighter power-law in the log-log plot (Figure 1A) than that seen in data for Brazil, USA and Japan (see Figure 1d of [15]). The clustering distribution (figure 3B) in Tanzania also differs from that in the USA and Japan, which both have a pronounced bump in the in-degree distribution (see Figure 4b of [15]).

We investigated a potential of the friend of a friend model as a local, self-organised mechanism for explaining the particular form of the network we found for Tanzania. The model gave a very good fit to both the degree distribution and the clustering distribution (figure 3B) with the parameters fit directly from the average degree ($n_q = 40$) and no other parameter tuning (if we can consider $q = 1$ as the parsimonious choice for this parameter). This indicates that the data is consistent with a process whereby new users are introduced to Twitter by a random friend, after which they then follow friends of that friend, with the average number of follows determined by the average degree. We are not proposing that it is precisely this model which determines every user's behaviour, but the model results do suggest that a similar self-organised mechanism of how new users are introduced to Twitter is a good starting point for thinking about the growth of social networks in sub-Saharan Africa.

One potential reason that the friend of a friend model fits so well is that the Tanzanian network has less external actors, large companies and sponsored accounts, working on it (as we saw in Figure 2). Large multinational companies than the networks of USA and Japan, in particular, skew the degree distribution away from one which is consistent with the model and a self-organised process. Further support for this claim can be found by noting that the Brazilian (with the data collected in 2012 in this case) clustering [15] is more similar to that of Tanzania (Figure 3B) than those of Japan and USA. The larger the influence of companies in a country the more a network deviates from being self-organised.

Studies of social media tend to focus on the Western world, often neglect sub-Saharan Africa countries. Our study has shown by looking closely at an African context, we can find societal structures that are much closer to those produced by idealised models. Such insights can be useful, not only in planning for development in sub-Saharan Africa countries, but also in helping Western countries learn how to create self-organised societal structures.

Availability of data and materials

All data analysed during this study are available on https://github.com/search?q=The-structure-of-Tanzania-Twitter-network&type=repositories.

Competing interests

The authors declare that they have no competing interests.

Funding

This research work received no specific funding.

Authors' contributions

WL and DJTS collected, analyzed and interpreted the Twitter data. EM and ISA analysed the data. All authors read and approved the final manuscript.

Acknowledgements

We thank Uppsala University for providing computer facilities for crowing Twitter data. We also thank Twitter for providing the API used to access data.

References

[1] Sadri AM, Hasan S, Ukkusuri SV, Lopez JES (2018b) Analysis of social interaction network properties and growth on Twitter. Soc Netw Anal Min 8(1):56. doi.org/10.1007/s13278-018-0533-y.

[2] Strogatz SH (2001) Exploring complex networks. nature, 410(6825): 268-276. doi.org/10.1038/35065725.

[3] Chen J, Dai M, Wen Z, Xi L (2014) Trapping on modular scale-free and small-world networks with multiple hubs. J Phys A 393:542–52. doi.org/10.1016/j.physa.2013.08.060.

[4] Watts DJ, Strogatz SH (1998) Collective dynamics of “small-world” networks. Nature, 393(6684), 440–442. doi:10.1038/30918

[5] Barabási AL,Albert R (1999) Emergence of scaling in random networks. science, 286(5439): 509-512. doi.org/10.1126/science.286.5439.509.

[6] Newman MEJ (2003) The structure and function of complex networks. SIAM Review 45:167-256. doi.org/10.1137/S003614450342480.

[7] Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B(2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp 29-42. doi.org/10.1145/1298306.1298311.

[8] Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Physical review E, 65(2): 026107. doi.org/10.1103/PhysRevE.65.026107

[9] Eikmeier N, Gleich DF (2017) Revisiting power-law distributions in spectra of real world networks. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 817-826. doi.org/10.1145/3097983.3098128.

[10] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: Structure and dynamics. Phys Rep 424:175–308. doi.org/10.1016/j.physrep.2005.10.009.

[11] Easley D, Kleinberg J (2010) Networks, crowds, and markets: Reasoning about a highly connected world, Vol 1. Cambridge: Cambridge university press. doi.org/10.1017/CBO9780511761942.

[12] Barthelemy M (2011) Spatial networks. Phys Rep 499:1–101. doi.org/10.1016/j.physrep.2010.11.002.

[13] Broido AD, Clauset A (2019) Scale-free networks are rare. Nature communications, 10(1): 1017. doi.org/10.1038/s41467-019-08746-5.

[14] Levens W, Szorkovszky A, Sumpter DJ (2022) Friend of a friend models of network growth. Royal Society Open Science.9(10): 221200. doi.org/10.1098/rsos.221200.

[15] Myers SA, Sharma A, Gupta P, Lin J (2014) Information network or social network? The structure of the Twitter follow graph. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 493-498. doi.org/10.1145/2567948.2576939.

[16] Poushter J, Bishop C, Chwe H (2018) Social media use continues to rise in developing countries but plateaus across developed ones. Pew research center, 22: 2-19.

[17] Tanzania members of parliament (2018) https://www.parliament.go.tz/mps-list

[18] Newman M (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5): 323–351. doi:10.1080/00107510500052444.

[19] Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM review, 51(4): 661-703. doi.org/10.1137/070710111.

[20] Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[21] Krapivsky PL, Redner S (2005) Network growth by copying. Physical Review E, 71(3): 036118. doi.org/10.1103/PhysRevE.71.036118.

[22] Bhat U, Krapivsky PL, Lambiotte R, Redner S (2016) Densification and structural transitions in networks that grow by node copying. Physical Review E, 94(6): 062302. doi.org/10.1103/PhysRevE.94.062302.

[23] Lambiotte R, Krapivsky PL, Bhat U, Redner S (2016) Structural transitions in densifying networks. Physical review letters, 117(21): 218301. doi.org/10.1103/PhysRevLett.117.218301.

[24] Twitter statistics: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/