Longitudinal Network Centrality Using Incomplete Data

Zachary C. Steinert-Threlkeld[1,*]

February 7, 2017

ABSTRACT

How does individuals' influence in a large social network change? Social scientists have difficulty answering this question because measuring influence requires frequent observations of a population of individuals' connections to each other, while sampling that social network removes information in a way that can bias inferences. This paper introduces a method to measure influence over time accurately from sampled network data. Ranking individuals by the sum of their connections' connections — *neighbor cumulative indegree centrality* — preserves the rank influence ordering that would be achieved in the presence of complete network data, lowering the barrier to measuring influence accurately. The paper then shows how to measure that variable changes each day, making it possible to analyze when and why an individual's influence in a network changes. This method is demonstrated and validated on 21 Twitter accounts in Bahrain and Egypt from early 2011. The paper then discusses how to use the method in domains such as voter mobilization and marketing.

Funding This work was supported by the United States Agency for International Development [(DF)#AID-OAA-A-12-00039]

Author's Note I would like to the anonymous reviewers and editorial staff at *Political Analysis* who provided insightful feedback. Participants of the Human Nature Group saw this paper from its infancy to its current form, and I am grateful for their patience. I am also fortunate to have had active audiences when this paper was presented at APSA and Sunbelt conferences. I would especially like to thank Lawrence Broz, James Fowler, Scott Guenther, Emilie Hafner-Burton, Will Hobbs, Alex Hughes, David Lake, David Lindsey, Mona Vakilifathi, and Barbara Walter for various forms of assistance. For replication material, see Steinert-Threlkeld (2016). Though I wish I could say otherwise, all remaining errors are mine.

[1]University of California, Los Angeles.
[*]To whom correspondence should be addressed: zst at luskin dot ucla dot edu

1. INTRODUCTION

Political scientists are increasingly interested in using network analysis to understand how individuals, institutions, and states influence each other over time (Lazer, Brewer, Christakis, Fowler & King 2009). Such work requires data on every connection each actor maintains with all other actors; these data are prohibitively costly to obtain when networks are large or change frequently. This cost is why existing time-series network analysis focuses on states at the year level (Dorussen & Ward 2008, Oatley, Winecoff, Pennock & Danzman 2013).

This paper introduces a statistic, *neighbor cumulative in-degree centrality (NCC)*, that allows for network time-series analysis of individuals at the daily level. NCC measures influence without data on every connection each actor maintains. Obviating the need for complete network data reduces research costs, allowing for daily network analysis of individuals. Moreover, NCC recovers an individual's influence that would be observed if complete data were available, and it outperforms the other measure, in-degree centrality, that is currently used with incomplete data.

NCC works best when the researcher can perform a breadth-first search - record all the connections for each individual being studied - and knows the number of connections each connection has. This situation is common for online social network data, as platforms such as Twitter and Instagram provide the number of connections each account has without the researcher having to manually download those connections. These data can also be obtained easily in surveys with a network component. For example, Karl-Dieter Opp and Christiane Gern surveyed participants in the 1989 Leipzig protests and asked if they had friends or co-workers who participated; if they had also asked respondents to rate those friends or co-workers on a popularity scale, the authors could have also determined if protestors are more likely to be influential in a network (high NCC) or not (low NCC) (Opp & Gern 1993).

NCC is also favorable when a researcher faces resource constraints. It is common for rate limits to slow the amount of information that can be downloaded from digital sources or lim-

ited funding to restrict the amount of data that can be gathered via in-person enumeration. Since it does not require complete network data, NCC uses much less data than common measures of influence such as eigenvector, PageRank, or closeness centrality. Only in small or stable networks such as a classroom, bill co-sponsorship, offices, or country alliances, among others, are other influence measures preferable.

The NCC measure is demonstrated in the context of activists and the Arab Spring. The influence (NCC ranking) of 21 Bahraini and Egyptian Twitter accounts is tracked over a three month period, as is those accounts' communication patterns. Models show that accounts which coordinate protests gain influence according to the NCC measure, while degree centrality influence suggests that the use of hashtags also matters. This result stands in contrast to work uses hashtag analysis to suggest the periphery of social networks drives protest mobilization (Barberá, Wang, Bonneau, Jost, Nagler, Tucker & González-Bailón 2015, Steinert-Threlkeld, Mocanu, Vespignani & Fowler 2015, Steinert-Threlkeld 2017).

Section 2 explains longitudinal analysis with NCC. Section 3 explains why to prefer node rankings instead of raw centrality scores, NCC to in-degree centrality, and under what situations NCC should be preferred to global centrality measures. Section 4 details a substantive application of the new measures: activism during the Arab Spring in Bahrain and Egypt. The main result of this analysis is that accounts which uses more hashtags become more influential based on degree centrality but not NCC, while both measures show that accounts become more influential when their messages coordinate protest. Section 5 provides detail on other applications of longitudinal NCC measurement; these methods can be used for scholars interested interested in identifying hidden influentials as well as voter mobilization, among other areas. Section 6 concludes.

2. NETWORK CENTRALITY, OVER TIME

2.1. *Network Centrality with Incomplete Data*

In network analyses, *centrality* refers to a set of statistics that attempt to measure which nodes are most influential, where the definition of influence varies according to the kind of network studied. ("Node" means the entity that forms the network under study. It could be an individual, a webpage, an internet router, an international organization, or a court case, for example. For the rest of this paper, "node" means individual, individual means node.) In this paper, Individual A is more influential than Individual B if the information he or she emits is seen by more people than that from Individual B. There are three main classes of centrality: betweenness, closeness, and degree-based. Each class of centrality measurement requires data on each node in the network (every website on the internet, every student in a school, or every nation in a trade network, for example) and the connections between those nodes (every link between webpages, every friend of each student, or the flow of trade between each country pair).

A node with a high betweenness centrality connects many nodes of a network; using this measure, the most important node is that which is on the most paths connecting any two nodes. Closeness centrality refers to the mean distance between one node and all other nodes; using this measure, the most important node is that which has the shortest average distance between itself and all other nodes.[1]

The most common centrality measures focus on the number of connections a node has to other nodes. The sum of these connections gives the degree of a node, and a node with higher degree is assumed to have more influence than one with lower degree. Measuring only the sum of connections of a node is called degree centrality or, in a directed network, indegree or outdegree centrality. Degree centrality is appealing because of its simplicity, but it does not give an indication of a node's position in the larger network: a node may have high degree

centrality, but if those with which it is connected have few connections, the node probably is not very important. Similarly, a node may not be connected to many other nodes, but if the nodes to which it is connected are themselves connected to many nodes, that node may be influential. A node can also be influential if it connects parts of a network that otherwise would not be connected.

Instead, a node's influence is also a function of the connections of that node's neighbors, its neighbors' neighbors, and so on. Many measures therefore take into account the importance of a node's neighbors to calculate a node's centrality, the idea being that an important node has neighbors that are also important. There are various ways to calculate these measures, some of the most common being eigenvector centrality, Katz centrality, PageRank, and k-core; see Newman (2010) for a mathematical explanation of these measures. For simplicity through the rest of the paper, I call these measures *global centrality measures*.

Eigenvector, Katz, PageRank, and k-core centrality require having data on every connection in a network. For example, studying how networks affect adolescent health in a high school would require knowing not just demographic data about each student but also who those students interact with; acquiring those data require large investments in time and money, and the cost multiplies with the duration of the study. As the network being studied grows, e.g. if one wants to study behaviors on Facebook or Twitter, calculating these centrality measures becomes exceedingly costly. Given this difficulty, degree centrality is the most common measure of centrality in large scale studies, especially those using social media datasets (Kwak, Lee, Park & Moon 2010, Garcia-Herranz, Moro, Cebrian, Christakis & Fowler 2014).

Degree centrality's appeal is therefore based on its ease of measurement, not its measurement validity. While it does correlate highly with global centrality measures (Bonner, Gilbert, Shi & Adamic 2008), that correlation masks heterogenous effects. Intuitively, a node with low degree could be connected to a node with very high degree, meaning whatever that node does could influence the larger network through its connection with the more

well-connected one; degree centrality does not capture this second-order effect, much less third or fourth-order ones. In a study using complete network data from Twitter, Facebook, Livejournal, and the American Physical Society, Pei et. al. (2014) find that global centrality measures, especially k-core centrality, better identify which nodes spread the most information (Pei, Muchnik, Andrade, Zheng & Makse 2014).[2] Degree centrality and PageRank are shown to create different rankings in a study of 41 million Twitter users from 2009 (Kwak et al. 2010). In other words, while the correlation between degree centrality and global centrality measures is high, the rank ordering correlation is much lower.

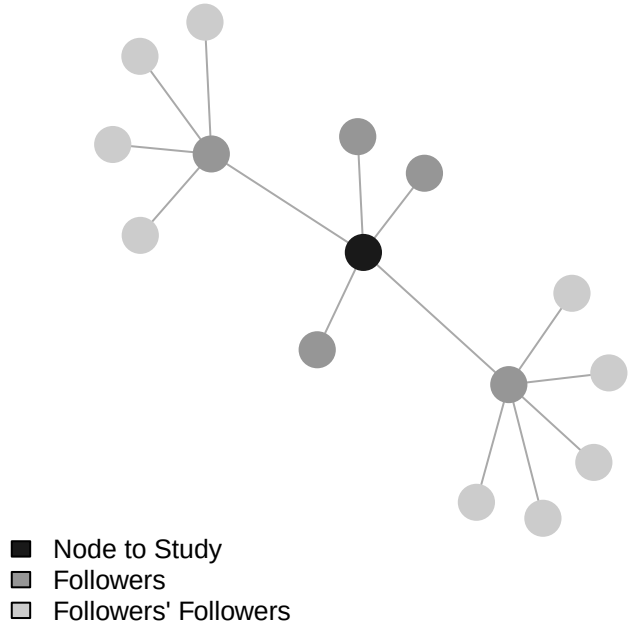
This paper introduces a measure that uses more data than degree centrality but less than global centrality measures. Specifically, a node's *neighbor cumulative indegree centrality (NCC)* is the sum of the indegree of the node's neighbor's. Formally:

$$NCC_i = \sum_{1}^{j} d_j \tag{1}$$

For each node i, the neighbor cumulative indegree centrality is the sum of the indegree centrality d_j for each neighbor j. This measure is first introduced in Pei et al. (2014) and has been used independently in Kim et al. (2015), though it does not appear to have yet gained widespread use. To the best of my knowledge, this paper is the first in political science to use it. Figure 1 presents an illustration of NCC.[3]

Simulations A series of simulations demonstrates that ranking by NCC instead of in-degree centrality more accurately recovers rankings based on eigenvector, PageRank, and closeness centrality. (Section 3 explains why rankings are preferred instead of raw scores.) For a series of networks ranging in size from 100 to 10,000 nodes, a power law degree distribution with a scaling exponent of 2.089, the scaling parameter found from three hours of streamed tweets, is used to assign connections between nodes, and each network contains ten times as many edges as nodes. The neighbor cumulative in-degree, eigenvector, PageRank, and closeness centrality of each node is then measured, and a node's influence is then determined

Figure 1: Degree Centrality = 5; Neighbor Cumulative Indegree Centrality = 9

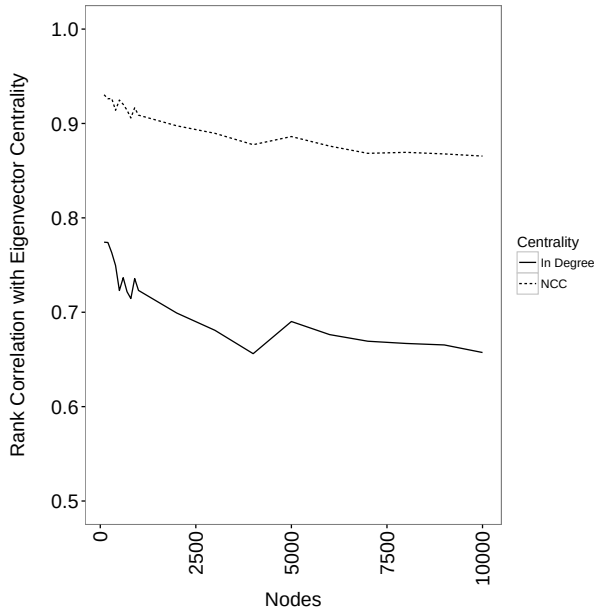


by its rank ordering based on each centrality score. For each network, a node's position in the NCC and in-degree rank orderings is compared to its position in the rank ordering based on eigenvector, PageRank, and closeness centrality. This comparison generates two bivariate graphs, one for NCC ranking and another for in-degree centrality. The correlation coefficients from those graphs are compared to each other for each network.

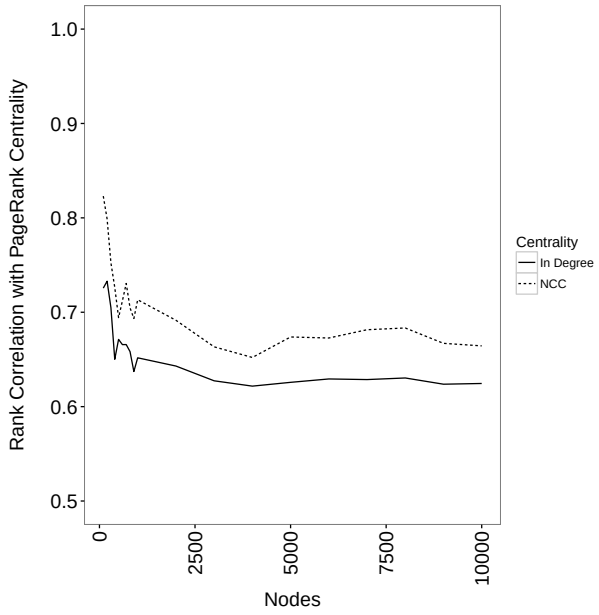
Figure 2 shows the result of this simulation. It shows that the rank ordering of nodes generated by neighbor cumulative indegree centrality preserves 70 to 90 percent of the rank ordering created by eigenvector, PageRank, and closeness centrality. Compared to in-degree centrality, this correlation represents a 7.89% improvement in rank correlation for PageRank centrality, 26.35% for closeness, and 26.77% for eigenvector. These results corroborate the empirical results of Pei et. al. (2014).

In many situations, however, the complete network is unavailable. I therefore also simulated networks, calculated global centrality measures, sampled nodes from the network, and

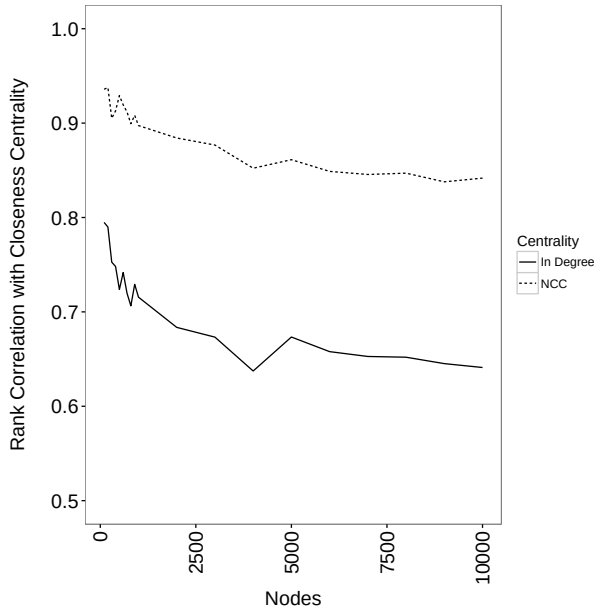
Figure 2: Neighbor Cumulative Indegree Centrality Better Measures Influence than Indegree Centrality



(a) Rank Ordering of Eigenvector Centrality



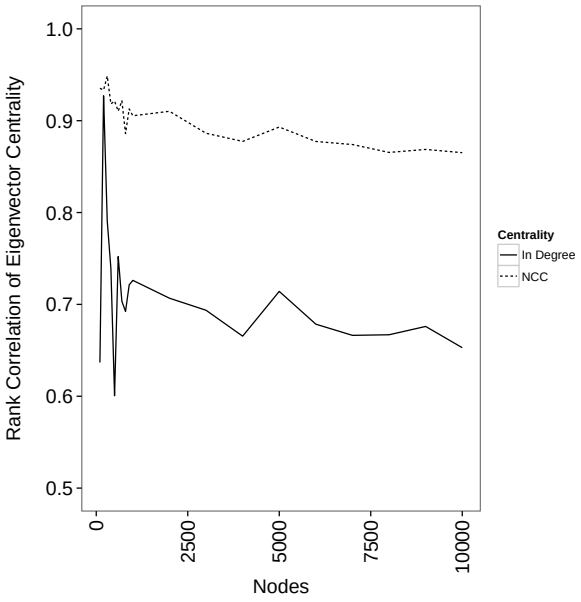
(b) Rank Ordering of PageRank Centrality



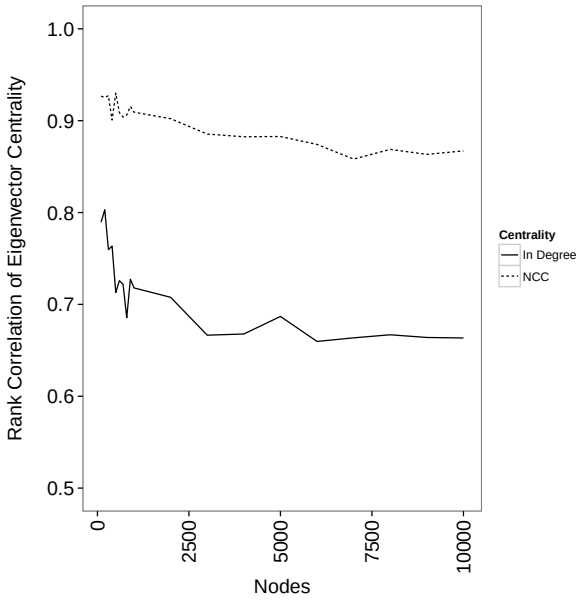
(c) Rank Ordering of Closeness Centrality

compared the rank correlation of NCC and in-degree centrality to those global centrality measures. Figure 3 shows these results comparing NCC and in-degree centrality to eigenvector centrality, and Section 1 of the Supplementary Materials shows the same for closeness and PageRank centrality. In sampled networks, NCC ranking continues to outperform in-degree centrality ranking.

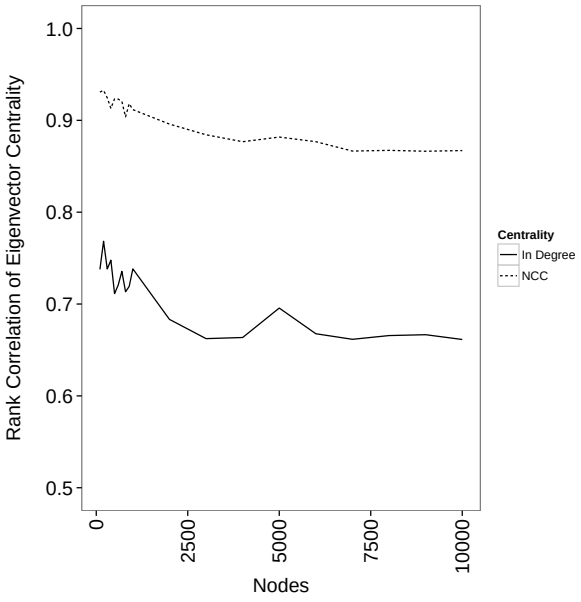
Figure 3: NCC and Eigenvector Ranking with Sampled Data



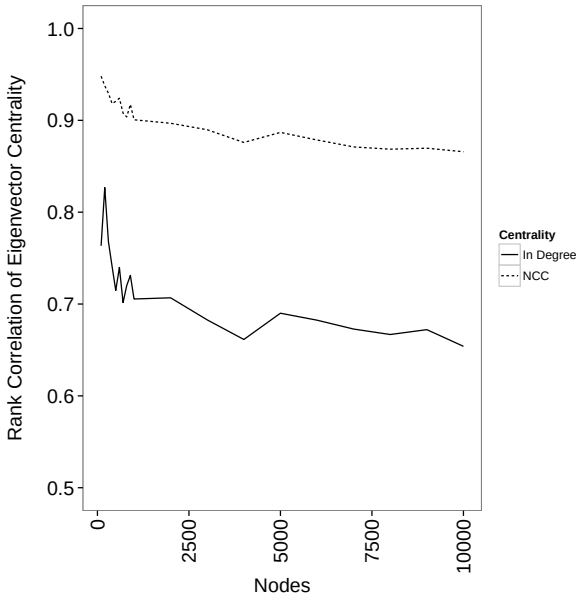
(a) Keep 13% of Nodes



(b) Keep 27% of Nodes



(a) Keep 38% of Nodes



(b) Keep 50% of Nodes

2.2. *Over Time*

To measure neighbor cumulative indegree centrality over time, a network needs to be measured at different points in time. If the network requires in-person measurement — surveying a school or canvassing a neighborhood, for example — that sampling procedure can be repeated and NCC measured a second time. If the network is measured digitally, such as via Twitter, the steps required to measure NCC longitudinally most likely differ from the the steps required to measure it initially. Because Twitter is one of, if not the, most common digital sources of network data, this section explains how to measure NCC over time using that platform.[4]

Twitter does not reveal when one user starts to follow the other, so a researcher only knows that a connection exists but not when it formed.[5] Two pieces of information from the REST API ameliorate this situation. First, the list of followers (or friends) that Twitter provides is sorted in reverse chronological order, meaning one knows the relative ordering of connection dates.[6] Second, the REST API provides the date when an account was created. These two pieces of information make it possible to accurately reconstruct when connections are formed.

Using the date followers join Twitter allows for the approximation of connection formation date, as shown in Table 1; because one does not know the precise date a connection forms, bounds around the actual date need to be created.[7] The lower bound of the bounds is calculated as follows: for each follower in a user's follower list, the earliest that that follower could have started following is the most recent Twitter joining date of all followers below that follower; this date is the lower bound of the estimate of the true connection forming date.

Estimating the upper bound on the connection formation date is more difficult; in fact, the upper bound itself has a lower and upper bound. The upper bound on the upper bound (UBUB) of the estimate of the connection formation date is the day the data were

downloaded, as it is theoretically possible an account's followers all started following that account earlier that day. The lower bound on the upper bound (LBUB) of the estimate of the connection date is the first Twitter joining date greater than that follower's Twitter joining date for the followers above that follower in the follower list. If no follower matches this criteria, the LBUB is the day the data were downloaded.

Table 1 clarifies this algorithm, and Section 2 of the Supplementary Materials provides pseudocode for it. Suppose User 1 has followers A, B, C, D, E, F, G, H, and I, with A the newest follower and I the oldest. Follower A joined Twitter on December 29^{th}, 2010 but could not have followed User 1 before 12.07.2013 because that is the most recent Twitter joining date of the nine followers. Follower C joined at the same time as A but could have started following User 1 as early as 06.20.2012 because the latest any of Followers C through I joined Twitter was that day. Follower G's earliest possible connection date is the same as the day it joined Twitter because neither of the two already existing followers joined Twitter after Follower G. These dates, the third column of Table1, are the lower bound of the estimate of the connection formation date.

The LBUB and UBUB of the estimate of the connection date is calculated as follows. Follower A's latest possible connection date is whatever day the follower list was downloaded, since Follower A is the newest follower of User 1; the same is true of Follower B because no subsequent follower (which is only Follower A) joined Twitter after Follower B. We can infer that Follower I connected to User 1 at least no later than 01.25.2011 because the first follower who connected with User 1 and had a Twitter joining date later than Follower I, Follower G, joined Twitter on 01.25.2011. Since Follower G could not follow User 1 before 01.25.2011, the LBUB for Follower I and User 1 is 01.25.2011; because we do not observe when Follower I actually started following User 1, the UBUB is the day the followers list was downloaded from Twitter. The same is true for Follower H; Follower H could not have connected to User 1 before Follower I, even though Follower H joined Twitter earlier, because Follower H is closer to the top of the follower list. The LBUB for Follower E and User 1 is 06.20.2012,

Table 1: Inferring Follower Relationship Formation

Follower*	Date Joined Twitter+	Lower Bound of Connection Date#	Lower Bound of Upper Bound of Connection Date#	Upper Bound of Upper Bound of Connection Date#
A	12.29.2010	12.07.2013	Day of API Request	Day of API Request
B	12.07.2013	12.07.2013	Day of API Request	Day of API Request
C	12.29.2010	06.20.2012	12.07.2013	Day of API Request
D	06.20.2012	06.20.2012	12.07.2013	Day of API Request
E	08.16.2009	03.15.2011	06.20.2012	Day of API Request
F	03.15.2011	03.15.2011	06.20.2012	Day of API Request
G	01.25.2011	01.25.2011	03.15.2011	Day of API Request
H	11.26.2008	08.16.2009	01.25.2011	Day of API Request
I	08.16.2009	08.16.2009	01.25.2011	Day of API Request

* From Twitter's GET followers/ids endpoint on the REST API.
+ From Twitter's GET users/lookup.
# Calculated by the researcher.

the first joining date of Followers A to D that is greater than Follower E's joining date of 08.16.2009.

The inability to establish a precise upper bound for the following date is theoretically problematic but pragmatically not. To return to Table 1, a researcher interested in the network of User 1 on 01.26.2011 can be certain that User 1 had at most two followers on that day. Theoretically, User 1 may have had 0 followers, if they all started following User 1 after 01.26.2011. But users gain followers over time; while bursty, users gaining all their followers on one day, which is what would be necessary for the upper bound of the upper bound of the confidence interval, is rare to nonexistent (Hutto, Yardi & Gilbert 2013, Antoniades & Dovrolis 2015, Myers & Leskovec 2014). Section 4.1 uses a dataset where users' true number of followers are known to show that the estimate accurately recovers the true number of followers. Meeder et. al. (2011) show that the estimate of the lower bound of the connection time accurately recovers the true connection time for celebrity accounts. Section 3 of the Supplementary Materials show that using the earliest latest date a connection forms quickly converges to the earliest date for accounts with hundreds of followers.

Meeder et. al. (2011) provide an analytic explanation of this process, and this paper builds on that work in three ways. First, it provides a method for estimating the upper bound of the follower connection date formation. Having a lower and upper bound for

follower connection dates allows for more precise estimation of connection formation, though the bounds approach each other as the number of followers increases. Second, Meeder et. al. (2011) work with celebrity accounts because they rapidly gain followers; the accounts in this sample show that this technique extends beyond celebrities. Third, the results show that measuring true changes in followers is accurate when combining the streaming and REST APIs, whereas Meeder et. al. (2011) use the REST API to crawl specific accounts. Since a large number of studies using Twitter, perhaps most, start with data from the streaming API, this paper provides a more realistic validation for estimating connection formation dates.

3. RANKING, NCC, AND WHEN TO USE RANKED NCC

Ranking nodes based on a centrality measure is preferable to using raw centrality measures, and ranking based on NCC is preferable to ranking on in-degree centrality, including in studies of offline social networks. NCC is to be preferred over global centrality measures when global network data are not available; global network data are rarely available because of cost.

3.1. *Ranking Instead of Raw Score*

There are two reasons to evaluate nodes by their rank instead of the absolute value of NCC. First, ranking individuals facilitates interpretation by controlling for unobserved heterogeneity. For example, individuals in the United States will have higher degree centrality and NCC than individuals in Suriname because the United States has more people; a user in Suriname with the same number of followers as one in the United States should therefore be more influential. Rank ordering at the country level, or whatever grouping makes the most sense for the research question, therefore acts as a fixed effect. Similarly, individuals in both

countries should see an increase in their degree centrality and NCC because of population growth.[8] Increases in degree centrality or NCC could erroneously be ascribed to a variable of interest when in fact the changes are a time effect. Rank ordering is therefore more likely to change as a result of a node's behaviors instead of unobservables. If using unranked NCC or in-degree centrality, individuals from a more populous setting will drive results.

Second, even if there is no concern about unobserved heterogeneity (all the observations are from the same school or country, for example), ranking has greater measurement validity than absolute values for most, perhaps all, social behaviors. Forbes publishes the 500 wealthiest individuals and largest corporations, not those worth $1 billion or with revenue over an arbitrary threshold. Olympic medals are given for the top three finishers, not everyone attaining a certain score or finishing below a certain time. Search engines returns pages in rank order of estimated relevance, not just those pages above a relevance threshold and certainly not randomly sorted. An A on an exam is less impressive if that is the modal grade than if a C is most common. In other words, social outcomes such as happiness, status, or influence, to name a few, derive from comparison to others, not to an abstract notion of those concepts (Brickman, Coates & Janoff-Bulman 1978, Veenhoven 1991, Adler, Epel, Castellazzo & Ickovics 2000). For researchers interested in influence in a network, relative influence (ranked NCC or in-degree centrality) should therefore also matter more than absolute influence (raw NCC or in-degree centrality).

Using ranking to evaluate nodes does not lead to different inferences than using absolute values. Sampled networks accurately recover the ranking of nodes based on degree, betweenness, and closeness centrality (Kim 2007). A canonical simulation of scale-free network growth, the Barabasi-Albert model, relies on new nodes knowing the degree of existing nodes (the "preferential attachment" mechanism) (Barabási & Albert 1999); it turns out that the same network can grow when new nodes only know the rank of existing nodes (Fortunato, Flammini & Menczer 2006). Even in gene regulatory networks, ranking by degree strongly correlates with complete centrality measures (Koschutzki & Schreiber 2008).

3.2. *NCC Instead of In-Degree Centrality*

Neighbor cumulative in-degree centrality has three advantages that compel its usage: it recovers influence rankings of global centrality measures better than in-degree centrality, does so at a significantly lower cost than those global centrality measures, and allows for centrality analysis on large offline networks.

First, the key benefit of NCC is that it recovers other centrality measures that require complete network data while using much less data. NCC works because it captures information on nodes up to two degrees away from the node for which NCC is calculated, incorporating much of information that global centrality measures incorporate while minimizing data requirements. The global centrality measures operate recursively, meaning they capture information on a node's 3rd, 4th, 5th, ... nth connections. While the contribution to importance of a node's third to n-th degree connections may matter, these far-away neighbors should have less of an effect than a node's immediate and second degree connections; empirically, this is the case (Christakis & Fowler 2012). On the other hand, in-degree centrality, as shown in the previous sections, generates misleading inferences about influence.

Another way to think about NCC is that it takes advantage of the power-law distribution of network degree that creates the friendship paradox (Feld 1991). Since a person's contacts will have more contacts, on average, than the original person, it is possible to monitor the emergence of behaviors by taking a sample of individuals and sampling the people to whom they are connected (Christakis & Fowler 2010, Garcia-Herranz et al. 2014).

Second, using much less data markedly lowers the cost of data collection. For example, Larson et. al (2016) collect the Twitter social network out to two degrees (the connections' connections) of 1,764 accounts from France, resulting in 199,126,639 additional nodes (111,618.07 connections per original account). The first-degree crawl this paper performs for the 21 activist accounts (discussed shortly) generates 90,863.52 connections per account. Gathering enough data to start analyzing network structure therefore requires at least 22.84%

more data; because this paper samples prominent accounts while Larson et al. sample more randomly, the computation differences are probably greater than 22.84%.

While Larson et al. (2016) do not undertake centrality analysis because it is not the focus of their research question, note that they would still have biased results because they do not have complete data. A comparison of sample strategies on four different networks finds that each sampling procedure requires a large network sample (over 50% of all nodes) before that sample's network characteristics converge to the full network's value (Lee, Kim & Jeong 2006). They could, however, calculate NCC, and because Twitter provides the number of followers for each account, calculating NCC from Twitter only requires a one-degree crawl.

Third, the need to collect data on all connections in a network in order to calculate centrality means that offline networks that have been studied are small. A canonical example is Zachary's karate club, where the social interactions of 34 members were observed over multiple years to understand why the club cleaved (Zachary 1977). A seven year study of dolphin social networks in a New Zealand fjord followed 83 dolphins (Lusseau, Schneider, Boisseau, Haase, Slooten & Dawson 2003). Scholars have made productive use of offline social network data for the 12,067 individuals in the Framingham Heart Study, though that study has received decades of generous institutional support that could not be replicated by an individual researcher (Christakis & Fowler 2007, Christakis & Fowler 2008, Fowler & Christakis 2008).

NCC increases the scale of network analysis that can be conducted without computers. For example, studies of social networks and political participation using surveys ask participants if they know people who also participated (McAdam 1986, Opp & Gern 1993) or observe the participation of individuals known to be connected to those treated by a survey instrument (Nickerson 2008) or online mobilization messages (Bond, Fariss, Jones, Kramer, Marlow, Settle & Fowler 2012). These studies do not, however, ask whether influence varies by how central individuals are in a network, as determining that centrality would have re-

quired each survey respondent to identify their friends, surveying those friends, asking those friends to name their friends, survey the friends' friends, and so on. Instead, if the survey asks each respondent to estimate the number of friends each friend has, the researcher can calculate NCC. This approach has been used in one study to optimize the spread of positive health behaviors, allowing researchers to identify influential individuals to treat (Kim, Hwong, Stafford, Hughes, O'Malley, Fowler & Christakis 2015). Since the data to calculate NCC can be gathered at the same time a survey is administered, centrality in larger offline networks can now be studied by smaller teams of researchers. Nickerson (2008), for example, surveyed 956 households, while Opp and Gern (1993) interviewed 1,300 individuals.

3.3. *When to Use*

Neighbor cumulative in-degree centrality is best suited for situations in which the researcher has a sampled network (which is most of the time) and can measure the number of connections a node's connections has.

Online social networks commonly provide the number of accounts a node follows or is followed by. For example, both Twitter and Instragram provide both sums as part of the user profile data. A researcher therefore only needs to download the user profile information of each account in a follower or following list in order to calculate the NCC of the accounts being studied. For example, the 21 accounts analyzed here have 1,908,134 followers, and those followers have a maximum of 506,821,726 followers. Calculating NCC for the 21 accounts does not require downloading 506,821,726 edges, however, as Twitter provides the number of followers as part of the profile information of each of the 1,908,134 first-degree followers. Recovering those nodes' centrality ranking that would be obtained with complete network data is therefore feasible with only a one-degree breadth first crawl.

Moreover, the lack of perfect correlation between NCC rank and rank based on complete centrality measures is due to change in rank for nodes with few connections; rank is more

stable for well-connected nodes than peripheral ones (Kim 2007, Cha, Haddadi, Benevenuto & Gummadi 2010). Because degree is power-law distribution, gaining 10 connections when one only has 10 will affect one's rank much more than gaining 10 when one has 1,000,000. For political scientists, this means that inferences based on well-connected groups of people - Congresspeople or members of the media, for example - will be more precise than for other groups. Precisely what "well-connected" means, however, is an open question. In this way, the use of NCC rank cannot circumvent a perennial issue: people on the margins of society are difficult to study, sometimes intentionally so.

When offline social network data are gathered, a researcher can ask an individual to estimate the number of friends his or her friends have. So long as those estimates are answered without bias, the resulting NCC rank of each respondent will approximate the rank that would be measured if the researcher counted the friends' friends him or herself. Collecting these data would require only one additional survey question or one more behavior to track if the researcher gathers data via participant observation. Relying on in-person data collection also makes it easier to study those who maintain few social connections.

If a researcher has complete network data (all nodes, all connections of those nodes, all those connections' connections, and so on), then it is preferable to use a global network centrality measure (eigenvector, PageRank, closeness, etc.) that takes advantage of the data. This situations rarely holds, however. Only in settings with few nodes or that can be closely monitored, such as a club, workplace, or school, will the entire network graph be observable. Even studies which use online social networks rarely observe second-degree effects of a treatment (see Bond et al. (2012) for an exception) or crawl the entire social graph (Larson et al. (2016), the most extensive recent crawl of Twitter, stops at friends of friends).

4. POLITICAL ENTREPRENEURS AND PROTEST MOBILIZATION

From Egypt and Bahrain, 42 activists representing five social movements were identified, 19 of whom were active on Twitter prior to each country's first protests. In Egypt, activists from the April 6th youth movement, the No Military Trials campaign, and the Anti-Sexual Harassment movement were chosen; in Bahrain, the human rights community and February 14th youth coalition were chosen, though only the human rights community was active on Twitter before the start of protests. The final 19 activists represent the three social movements in Egypt and Bahrain's human rights community. These movements were chosen because they were active before, during, and after each country's main protest period, and individuals accounts were identified in collaboration with a colleague at a British university; for more detail on the movements and accounts, see Fowler and Steinert-Threlkeld (2016) for more detail. Two Bahraini government accounts were also identified and collected, raising the final number of accounts under analysis to 21.

Their position in the larger Twitter social network and their behaviors are observed from January 11th, 2011 to April 5th, 2011. Measuring NCC requires working with Twitter's REST API. I also purchased these accounts' tweets from early 2011 to confirm the accuracy of NCC measure; each tweet provides data on how many followers an account has at the time it is created, providing a ground-truth to which to compare the followers estimate (Shulman 2011). See Section 5 of the Supplementary Material for a discussion of these accounts, why they were chosen, the Arab Spring, and more information on acquiring their data.

4.1. *Reconstructing Daily Network Change*

This section demonstrates that the procedure in Section 2.2 accurately measures the true number of followers and reveals changing network structure. The results are presented using

the lower bound of the estimate of the connection date (column 3 from Table 1), and Section 3 of the Supplementary Materials show that results do not change if using the lower bound of the upper bound of the estimate of the connection date (column 4 from Table 1).

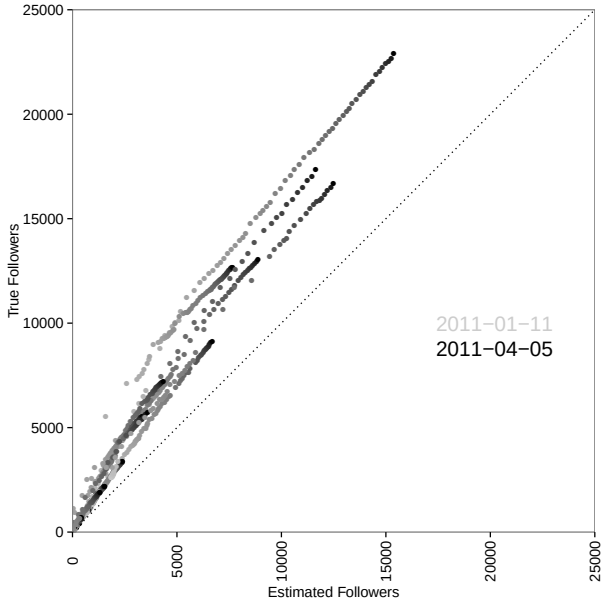
There are two ways to measure a user's change in followers over time: either observe that user in real-time (with the streaming API) while frequently downloading their followers' list (via the REST API), or estimate, later and indirectly, that change. The former is most precise but requires that the researcher knows which accounts he or she is interested in before an account is observed for a study. Estimating the change indirectly, through the REST API, is therefore how most longitudinal analyses will proceed. This section demonstrates that estimating indirectly the change accurately recovers the true number of followers and can show daily change in network structure, substantiating the methodology explained in Section 2.2.

Figure 4a shows that the *post hoc* estimated number of followers linearly predicts the true number of followers. The estimated number of followers underpredicts the true number because users can stop following an account or delete their account, the followers list was downloaded after the period of study, and Twitter removes users from the followers list once they stop following an account.

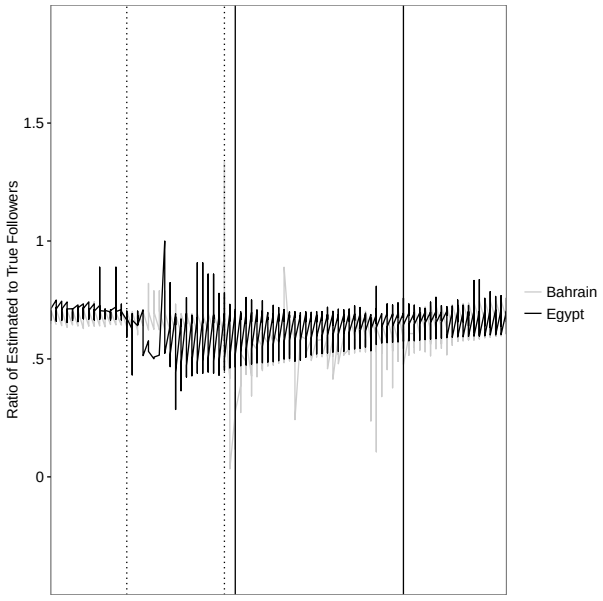
In Figure 4a, accounts are shaded from light to dark gray based on how close to April 5th, 2011 they are.[9] The estimated number of followers explains 98.09% of the variance in the number of true followers, with half of the remaining variance explained by group fixed effects; both these estimates are based on a linear model not shown here.[10] The residual increases as a function of the estimated number of followers, but this heteroskedasticity is constant as a percentage of an account's followers.

Figure 4b shows that the estimated number of followers is usually 67.53% of the true number of followers. This relationship holds whether or not the results are pooled by country; aggregating observations by group does not change the trends. The dashed lines correspond to the start and end of protests in Egypt, the solid in Bahrain. The *post hoc* measure performs

Figure 4: Verification Against Ground Truth Data



(a) Accuracy of Approximation of Followers' Daily Change



(b) Accuracy by Country, Day

less consistently, though does not appear biased, during these protest periods, suggesting that the measure may perform less well when the number of followers fluctuates rapidly. Overall, the *post hoc* measure of followers consistently approximates the true measure, suggesting it can be used when the true number of friends is not observable.

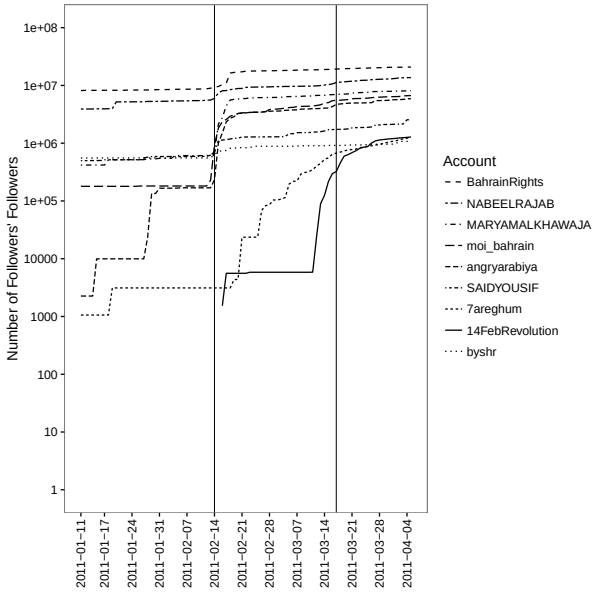
4.2. *Daily Changes in NCC*

To measure neighbor cumulative indegree centrality, the user ID of each of the 21 seed account's followers was downloaded from Twitter's GET users/ids endpoint, returning 4,229,373 results containing 1,908,134 unique followers. Each user ID was then submitted to Twitter's GET users/lookup endpoint, providing data such as when the user joined Twitter, their self-reported location, their default language, and how many tweets they have authored. These first-degree followers themselves have 506,821,726 followers. Since downloading the second-degree connections would require six months, and weeks more to download metadata for each ID, data on second-degree connections were not acquired.

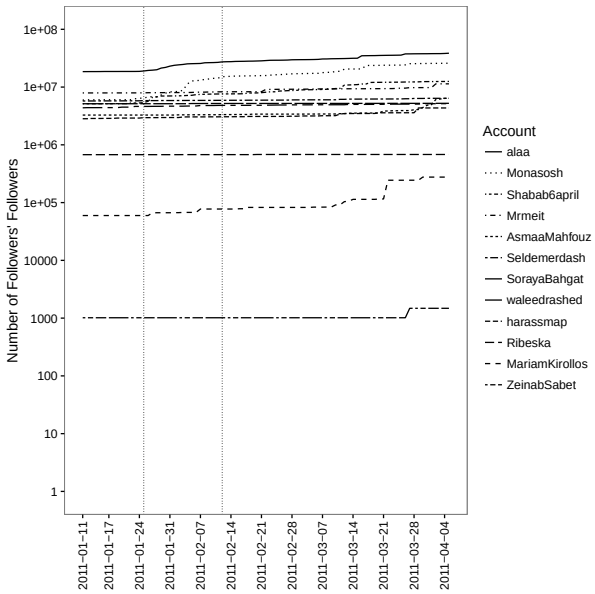
Figure 5 presents the change in neighbor cumulative indegree centrality over three months in Bahrain and Egypt. The first vertical line represents the start of protests, the second the end. Each country's legend is ordered from highest to lowest values of NCC at the end of the period. Color figures are in Section 4 of the Supplementary Materials.

A few results emerge from Figure 5. In both countries, relative influence is stable: the rank ordering of NCC on January 11th, 2011 looks very similar to that on April 5th, 2011. Even though every account except for @Ribeska gains NCC, very few gain influence at a quicker rate than their peers. In Bahrain, a notable change is @angryarabiya, who moves from second least influential to fifth most; that account belongs to the daughter of Nabeel Rajab (@NABEELRAJAB), a human rights advocate who led — he is now imprisoned — the Bahrain Center for Human Rights (@BahrainRights). The Ministry of the Interior's account, @moi_bahrain, is the fourth most influential at the end of the study, an increase of two spots.

Figure 5: Reconstructed Temporal Change in Influence



(a) Bahrain



(b) Egypt

@byshr, the account of the Bahrain Youth Society for Human Rights, experiences the steepest decline, moving from third to last. Egypt's relative ordering is more stable. @Shabab6april experiences the greatest change in NCC, moving from fifth to third. @monasosh experiences a large increase in absolute influence, but she only moves from the third to second most influential account.

Both countries' accounts also experience the greatest changes in NCC and rank ordering around their protest periods. Each country's users start to gain influence days before the start of protests. Most continue to gain influence during the protest period, and some stabilize after while others continue to gain influence.

Finally, comparing the NCC across Bahrain and Egypt reveals differing network properties. The Bahrain accounts start and end with lower average NCC than the Egyptian ones. Egypt, on the other hand, has higher variance in NCC. The three least influential Egyptians accounts, the relative ranking of which do not change, are accounts for individuals associated with the Anti-Sexual Harassment movement. That movement has been more peripheral to Egyptian politics than those sampled in Bahrain. Excluding those three, the Egyptian accounts have greater influence and lower variance than the Bahraini ones. Why countries' networks have different structural properties is outside of this paper's scope, but has started to receive some attention (Zeitzoff, Kelly & Lotan 2015).

4.3. *Individual Behavior and Changes in NCC*

The temporal change of neighbor cumulative degree centrality can be combined with accounts' tweeting behavior to analyze if certain patterns of behavior change an account's influence in a network.

Table 2 reveals that the effect of individual behaviors varies depending on which measure of centrality is used. Table 2 shows the results from regressing measures of the 21 accounts' position in their Twitter network on measures of their behavior and account fixed effects.

The dependent variable is the rank of an account on a day, depending on whether the measure is degree centrality (column 1) or neighbor cumulative degree centrality (column 2). The independent variables are the number of tweets from an account, the number of tweets with hashtags, the number of tweets that mention another user, the number of tweets that coordinate protest activity, and account fixed effects.[11] Because the dependent variable is a ranking, a negative coefficient means that an increase of that variable corresponds to increased influence.

Table 2 shows that more tweets with hashtags are not associated with greater influence. The results in Table 2 show that a model of influence which relies on degree centrality will suggest that an account which tweets more using hashtags will have a lower ranking than if it did not. While some work has argued that the best way to increase one's influence on Twitter is to use hashtags to make one's tweets part of a larger conversation (Kwak et al. 2010, Bruns & Burgess 2011, Gonzalez-Bailon, Borge-Holthoefer & Moreno 2013), this finding corroborates other researchers who find that specializing in a particular topic on Twitter is how accounts gain influence (Cha et al. 2010). While hashtags may decrease one's ranking based on the number of followers (Column 1), it does not appear to do so based on the followers those followers have (Column 2). In other words, using hashtags may cause an account to gain followers but not at a greater rate than other individuals in the network. Moreover, those followers do not have many followers, causing no change in influence as measured by NCC.

Both models find that more tweets coordinating protests leads to an account being ranked more highly. On the other hand, the only variable which leads to an increase of NCC rank (column 2) is the number of tweets about protest coordination. This result is in line with other work that has found that user's influence rank, measured by retweets and mentions, increases as they specialize in tweeting about one topic (Cha et al. 2010). Note as well that a model of NCC Rank fits better than a model of Degree Centrality Rank.

Table 2: Individual Correlates of Structural Position

	Degree Centrality $\text{Rank}_{i,t}$	NCC $\text{Rank}_{i,t}$
	(1)	(2)
$\text{Tweets}_{i,t-1}$	−0.003	0.004
	(0.006)	(0.007)
$\text{Hashtags}_{i,t-1}$	0.019	0.009
	(0.006)	(0.007)
$\text{Mentions}_{i,t-1}$	−0.004	−0.009
	(0.007)	(0.008)
$\text{Coordination}_{i,t-1}$	−0.034	−0.030
	(0.011)	(0.013)
Account FE	Yes	Yes
N	1,080	1,080
AIC	2261.125	1831.805

All models are ordered logit with a ranked dependent variable.
A negative sign means a node becomes more influential.
Model 1 with a lagged dependent variable fails to converge.
Model 2 with a lagged dependent variable has the same results.

5. OTHER APPLICATIONS

This section details other domains in which longitudinal neighbor cumulative in-degree centrality is useful.

5.1. *Hidden Influentials*

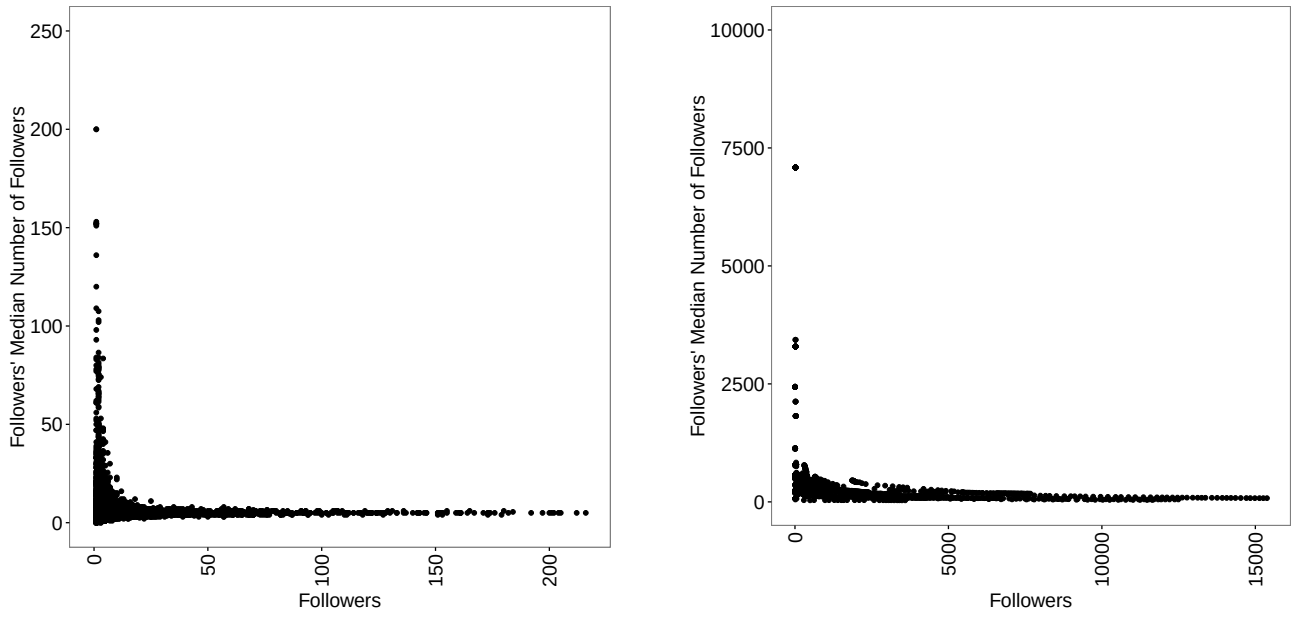
Network studies often are interested in identifying which nodes facilitate diffusion. While it is common to analyze highly central nodes, recent work on protest diffusion suggests that accounts with low out-degree but high in-degree may also be influential; these accounts are called "hidden influentials" and refer to accounts that global centrality measures may miss(Gonzalez-Bailon, Borge-Holthoefer & Moreno 2013). A slight modification of the NCC measure suggests an alternate method of finding hidden influentials.

Instead of taking the sum of neighbors' indegree centrality, the median of the neighbors' indegree identifies accounts whose followers have many followers. NCC favors accounts with many followers, with some weight assigned to how popular those followers are; there thus exists a strong positive relationship between the number of followers and the sum of the followers' followers. Taking the median of the followers' followers emphasizes accounts with few followers but whose followers' followers have many connections; it is preferred to the average so that one or two very popular followers does not bias results. The accounts with a high median number of followers' followers may be hidden influentials.

Figure 6a shows the simulated distribution of the median NCC against the distribution of followers; these data are the same used in Figure 2. Figure 6b is the same but on the data from the 21 accounts. In both cases, there is a clear decaying relationship between the number of followers and the median NCC. This decaying relationship makes sense, as most individuals in a network have few connections while a few have very many (Feld 1991). In

the simulated and actual data, however, there are some accounts that have few followers but whose followers have very many followers. These hidden influentials are the data points crawling up the y-axis near x equals 0. Because these data are a for a directed network, these accounts are those who are followed by accounts with many followers even though they themselves are not followed by many.

Figure 6: Median Neighbor Indegree Centrality and Hidden Influentials



(a) Simulation

(b) 21 Accounts

Using median NCC to identify accounts may reveal nodes in a network which help products diffuse or campaign messages resonate. Marketers understand that diffusion on a network is likely to come from those with many followers, but which of those central individuals will cause diffusion is very hard to predict. This apparent randomness means marketers have to target all "influencers", a costly proposition (Bakshy, Hofman, Watts & Mason 2011). Instead, a better approach may be to identify and target those accounts that the influencers follow, as they will be less expensive. Targeting these hidden influentials may be a more attractive option than focusing on the mass of individuals whose weak links to each other otherwise spread information about products (Watts & Dodds 2007, Bakshy, Marlow, Rosenn

& Adamic 2012).

5.2. *American Politics*

Since President Obama's 2008 election, scholars have realized the value large datasets have for political scientists (Nickerson & Rogers 2014). Studies which use network concepts to measure behaviors of interest to American politics have traditionally relied on cross-section surveys, and I am aware of no work which uses longitudinal network analysis. The following sections briefly discuss possible applications of NCC.

A large literature examines the conditions under which individuals mobilize to vote; for reviews of it, see Blais (2006) and Jacobson (2015). Part of that literature focuses on how individuals' social connections affect their decision making, with a heavy use of cross-section surveys and field experiments to make causal claims (Huckfeldt & Sprague 1987, Lake & Huckfeldt 1998). Work that does incorporate a temporal component focuses on political institutions like Congress or the Supreme Court because they contain few individuals and make data collection relatively easy (Fowler, Johnson, Spriggs II, Jeon & Wahlbeck 2007, Rogowski & Sinclair 2012). Scholars have not, however, been able to study voters in their networks over time. Does an individual's network position change in response to his or her political beliefs? Is one more likely to vote if someone central to their network does so? If an individual's friend expresses a differing political opinion, does the centrality of that friend affect the individual's likelihood to change opinion? Do elections affect the structure of one's friendship networks? If so, does the effect vary for local, state, and presidential elections? These questions can start to be answered with the methods presented in this paper.

Political parties target voters in order to persuade them to support their candidate, and the methods developed in this paper may help them identify influential individuals to target. Prior to campaigns' ability to use large datasets to target specific individuals (Hersh & Schaffner 2013, Nickerson & Rogers 2014), campaigns would canvass large groups of people,

hoping to create a "ripple effect of social interaction" in their favor (Sprague & Huckfeldt 1992, pg. 77). Parties vary their contact based on supra-individual characteristics, such as district or state competitiveness, and have done so since at least 1956 (Panagopoulos & Francia 2009). The methods developed here, however, could allow a campaign to distinguish influential core supporters from non-influential ones or find influential individuals socially near a campaign's core supporters (Holbrook & McClurg 2005). The NCC measure can also identify which peripheral individuals are influential, letting a campaign focus more efficiently use its resources to persuade them (Chen & Reeves 2011). The ability to observe communities evolve can alert campaigns to groups of people who have followed their candidate as well as ideologically far ones; assuming those individuals have not decided who to support, targeting them before the competition does would be valuable (Huckfeldt, Mendez & Osborn 2004).

6. CONCLUSION

This paper joins a growing body of longitudinal network analysis in political science, but it is the first, as far I am aware, to analyze individuals at the daily level. Longitudinal network analysis has been used to understand the Great Recession (Oatley et al. 2013), the effect of international organization of conflict (Hafner-Burton & Montgomery 2006, Dorussen & Ward 2008), the relationship between trade and conflict (Lupu & Traag 2013), and jurisprudence at the European Court of Human Rights (Lupu & Voeten 2012). These studies analyze cases, institutions, or states as their relationships change every year. The population of each is much smaller than the population of people, and focusing on annual change lowers the cost of data collection. NCC allows the researcher to analyze changes in populations heretofore too large to study, and the lower cost of calculating it facilitates the measurement of daily changes.

While multiple online social networks exist that could provide data, this paper focuses on Twitter. Twitter's global reach, large user base, and data openness make it a common

platform for large-scale studies of human behavior. With over 300 million accounts creating 500 million messages per day, it is one of the largest online social networks. Its data are also relatively easy to access, compared to other platforms. While other social media platforms and websites, such as reddit or Instagram, also have easily accessible data, none are as general purpose as Twitter. Though Twitter is the preferred platform for analyses of networks through social media, analyzes of network structure with its data are difficult because of how the platform provides data to researchers. Data provided as a streaming sample make structure difficult to see, while Twitter limits how often one can download data on connections between individuals. This paper's methods work within Twitter's limits.

While neighbor cumulative indegree centrality captures rank ordering that would be obtained with complete network data, it may still be preferable to have information on more than first-degree connections; for example, one can start topographic analysis with data on connections' connections (Larson, Nagler, Ronen & Tucker 2016). In practice, such information is very costly to obtain. Because the number of connections in a network expands exponentially while Twitter's rate limits are fixed, computing time increases supralinearly. For the 21 users in this study, their 1,908,134 followers have 506,821,726 followers; at 60 requests per hour returning a maximum of 5,000 followers per request, one computer connection would need just over 70 days to download the list of 2nd-degree followers. Assuming 45% of those are unique (the percentage from the crawl of followers for this paper), one computer would require almost 132 days to download data on each unique user. While this number is probably an overestimate, since some of the second-degree followers may have been followers of one of the other 21 accounts, the rate at which the download time increases as a function of degrees from a seed node is unknown. A complete crawl of Twitter conducted in July 2012 used two machines that could make 20,000 requests per hour, two that could make 100,000, and 550 machines using the normal rate limits; this crawl required 4 months and 4 days (Gabielkov, Rao & Legout 2014). The four machines with higher rate limits were whitelisted, a now defunct practice by which Twitter gave certain machines preferential ac-

cess to their data. A similar crawl without whitelisted machines would therefore take about double the time, according to those authors' estimates.

The main barrier presently facing researchers is therefore programming rate limits. Future work should explore how to approximate neighbor cumulative indegree centrality without having to sample all of a node's followers. Because of the way Twitter returns data, the approximation would need to work with the newest followers of an account.

This paper has also only treated one direction of an asymmetric network, treating accounts as emitters of information. But individuals also consume information, and the consumption network should change over time as well. The symmetric network — where each connection represents mutual following — will also reveal patterns about more intimate types of relationships. How these networks change over time remains an open question.

Finally, these methods can be used to study offline networks. It is common for studies of networks and political behavior to administer surveys and ask respondents to name their friends (McAdam 1986, Opp & Gern 1993). Modifying this approach, a researcher could ask those the respondent names how many friends they have or even ask the respondent how many friends she or he thinks each of the friends has. This information would be enough to generate NCC scores for the original respondents. Generating the NCC from offline data allows researchers who do not use online social network datasets or who are interested in samples of individuals not on these networks to also approximate centrality when full network data are not available.

REFERENCES

Adler, Nancy E., Elissa S. Epel, Grace Castellazzo & Jeannette R. Ickovics. 2000. "Relationship of Subjective and Objective Social Status With Psychological and Physiological Functioning: Preliminary data in Healthy White Women." *Health Psychology* 19(6):586–592.

Antoniades, Demetris & Constantine Dovrolis. 2015. "Co-evolutionary dynamics in social networks: A case study of Twitter." *Computational Social Networks* 2(1):1–21.

Bakshy, Eytan, Cameron Marlow, Itamar Rosenn & Lada Adamic. 2012. The Role of Social Networks in Information Diffusion. In *International World Wide Web Conference*. Lyon: ACM pp. 519–528.

Bakshy, Eytan, Jake M. Hofman, Duncan J. Watts & Winter A. Mason. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. Hong Kong: ACM Press pp. 65–74.

Barabási, Albert-László & Reka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286(October):509–513.

Barberá, Pablo, Ning Wang, Richard Bonneau, John T. Jost, Jonathan Nagler, Joshua Tucker & Sandra González-Bailón. 2015. "The Critical Periphery in the Growth of Social Protests." *PloS ONE* 10(11):1–15.

Blais, André. 2006. "What Affects Voter Turnout?" *Annual Review of Political Science* 9:111–125.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D.I. Kramer, Cameron Marlow, Jaime E. Settle & James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295–8.

Bonner, Matthew, Anna C. Gilbert, Xiaolin Shi & Lada Adamic. 2008. The Very Small World of the Well-Connected. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*. Pittsburgh: pp. 61–70.

Brickman, Philip, Dan Coates & Ronnie Janoff-Bulman. 1978. "Lottery winners and accident victims: is happiness relative?" *Journal of Personality and Social Psychology* 36(8):917–927.

Bruns, Axel & Jean E Burgess. 2011. The Use of Twitter Hashtags in the Formation of Ad Hoc Publics. In *6th European Consortium for Political Research Conferenc*. Number August Reykjavik: pp. 1–10.

Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto & Krishna P Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Washington D.C.: AAAI pp. 10–17.

Chen, L. J. & a. Reeves. 2011. "Turning Out the Base or Appealing to the Periphery? An Analysis of County-Level Candidate Appearances in the 2008 Presidential Campaign." *American Politics Research* 39:534–556.

Christakis, Nicholas A & James H Fowler. 2007. "The spread of obesity in a large social network over 32 years." *The New England Journal of Medicine* 357(4):370–9.

Christakis, Nicholas A. & James H. Fowler. 2008. "The Collective Dynamics of Smoking in a Large Social Network." *New England Journal of Medicine* 358(21):2249–2258.

Christakis, Nicholas A & James H Fowler. 2010. "Social network sensors for early detection of contagious outbreaks." *PloS ONE* 5(9):e12948.

Christakis, Nicholas A. & James H. Fowler. 2012. "Social contagion theory: examining dynamic social networks and human behavior." *Statistics in Medicine* 32(4):556–577.

Dorussen, Han & Hugh Ward. 2008. "Intergovernmental Organizations and the Kantian Peace: A Network Perspective." *The Journal of Conflict Resolution* 52(2):189–212.

Feld, Scott L. 1991. "Why Your Friends Have More Friends than You Do." *American Journal of Sociology* 96(6):1464–1477.

Fortunato, Santo, Alessandro Flammini & Filippo Menczer. 2006. "Scale-free network growth by ranking." *Physical Review Letters* 96(21):1–4.

Fowler, James H. & Nicholas A. Christakis. 2008. "Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study." *BMJ* 337(a2338):1–9.
URL: *http://www.bmj.com/cgi/doi/10.1136/bmj.a2338*

Fowler, James H., Timothy R. Johnson, James F. Spriggs II, Sangick Jeon & Paul J. Wahlbeck. 2007. "Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court." *Political Analysis* 15(3):324–346.

Fowler, James & Zachary C. Steinert-Threlkeld. 2016. Online and Offline Activism in Egypt and Bahrain. Technical report United States Agency for International Development.
URL: *http://www.iie.org/en/Research-and-Publications/Publications-and-Reports/IIE-Bookstore/DFG-UCSD-Publication#.V-MIM5MrKqA*

Gabielkov, Maksym, Ashwin Rao & Arnauld Legout. 2014. Studying Social Networks at Scale : Macroscopic Anatomy of the Twitter Social Graph. In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*. Austin: ACM Press pp. 277–288.

Garcia-Herranz, Manuel, Esteban Moro, Manuel Cebrian, Nicholas A. Christakis & James H. Fowler. 2014. "Using Friends as Sensors to Detect Global-Scale Contagious Outbreaks." *PloS ONE* 9(4):e92413.

Gonzalez-Bailon, Sandra, Javier Borge-Holthoefer & Yamir Moreno. 2013. "Broadcasters and Hidden Influentials in Online Protest Diffusion." *American Behavioral Scientist* 57(7):943–965.

Hafner-Burton, Emilie M. & Alexander H. Montgomery. 2006. "Power Positions: International Organizations, Social Networks, and Conflict." *Journal of Conflict Resolution* 50(1):3–27.

Hersh, Eitan D. & Brian F. Schaffner. 2013. "Targeted Campaign Appeals and the Value of Ambiguity." *The Journal of Politics* 75(02):520–534.

Holbrook, Thomas M. & Scott D. McClurg. 2005. "The mobilization of core supporters: Campaigns, Turnout, and Electoral Composition in United States Presidential Elections." *American Journal of Political Science* 49(4):689–703.

Huckfeldt, Robert, Jeanette Morehouse Mendez & Tracy Osborn. 2004. "Disagreement, Ambivalence, and Engagement: The Political Consequences of Heterogenous Networks." *Political Psychology* 25(1):65–95.

Huckfeldt, Robert & John Sprague. 1987. "Networks in Context: The Social Flow of Political Information." *American Political Science Review* 81(4):1197–1216.

Hutto, C.J., Sarita Yardi & Eric Gilbert. 2013. A Longitudinal Study of Follow Predictors on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris: ACM Press pp. 821–830.

Jacobson, Gary C. 2015. "How Do Campaigns Matter?" *Annual Review of Political Science* 18:31–47.

Kim, David A., Alison R. Hwong, Derek Stafford, D. Alex Hughes, A. James O'Malley, James H. Fowler & Nicholas A. Christakis. 2015. "Social network targeting to maximise population behaviour change: A cluster randomised controlled trial." *The Lancet* 386(9989):145–153.

Kim, So Young. 2007. "Openness, External Risk, and Volatility: Implications for the Compensation Hypothesis." *International Organization* 61(01):181–216.
URL: *http://www.journals.cambridge.org/abstract_S0020818307070051*

Koschutzki, Dirk & Falk Schreiber. 2008. "Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks." *Gene Regulation and Systems Biology* 2008(2):193–201.

Kwak, Haewoon, Changhyun Lee, Hosung Park & Sue Moon. 2010. What is Twitter, a Social Network or a News Media ? In *International World Wide Conference*. pp. 591–600.

Lake, Ronald La Due & Robert Huckfeldt. 1998. "Social Capital, Social Networks, and Political Participation." *Political Psychology* 19(3):567–584.

Larson, Jennifer M., Jonathan Nagler, Jonathan Ronen & Joshua A. Tucker. 2016. "Social Networks and Protest Participation: Evidence from 130 Million Twitter Users.".

Lazer, David, Devon Brewer, Nicholas Christakis, James Fowler & Gary King. 2009. "Life in the network: the coming age of computational social science." *Science* 323(5915):721–723.

Lee, Sang, Pan-Jun Kim & Hawoong Jeong. 2006. "Statistical properties of sampled networks." *Physical Review E* 73(1):016102.

Lupu, Y & E Voeten. 2012. "Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights." *British Journal of Political Science* 42(02):413–439.

Lupu, Yonatan & Vincent A. Traag. 2013. "Trading Communities, the Networked Structure of International Relations, and the Kantian Peace." *The Journal of Conflict Resolution* 57(6):1011–1042.

Lusseau, David, Karsten Schneider, Oliver J. Boisseau, Patti Haase, Elisabeth Slooten & Steve M. Dawson. 2003. "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait?" *Behavioral Ecology and Sociobiology* 54(4):396–405.

McAdam, Doug. 1986. "Recruitment to High-Risk Activism: The Case of Freedom Summer." *American Journal of Sociology* 92(1):64–90.

Meeder, Brendan, Brian Karrer, Christian Borgs, R Ravi & Jennifer Chayes. 2011. We Know Who You Followed Last Summer: Inferring Social Link Creation Times in Twitter. In *Proceedings of the 20th International Conference on World Wide Web Pages*. pp. 517–526.

Myers, Seth A. & Jure Leskovec. 2014. The bursty dynamics of the Twitter information network. In *Proceedings of the 23rd International Conference on the World Wide Web*. Seoul: ACM Press pp. 913–924.

Newman, M.E.J. 2010. *Networks: An Introduction*. Oxford: Oxford University Press.

Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102(01):49–57.

Nickerson, David W. & Todd Rogers. 2014. "Political Campaigns and Big Data." *Journal of Economic Perspectives* 28(2):51–74.

Oatley, Thomas, W. Kindred Winecoff, Andrew Pennock & Sarah Bauerle Danzman. 2013. "The Political Economy of Global Finance: A Network Model." *Perspectives on Politics* 11(01):133–153.

Opp, Karl-Dieter & Christiane Gern. 1993. "Dissident Groups, Personal Networks, and Spontaneous Cooperation: The East German Revolution of 1989." *American Sociological Review* 58(5):659–680.

Panagopoulos, Costas & Peter L. Francia. 2009. "Grassroots Mobilization in the 2008 Presidential Election." *Journal of Political Marketing* 8(4):315–333.

Pei, Sen, Lev Muchnik, José S Andrade, Zhiming Zheng & Hernán a Makse. 2014. "Searching for superspreaders of information in real-world social media." *Scientific reports* 4(c):5547.

Rogowski, J. C. & B. Sinclair. 2012. "Estimating the Causal Effects of Social Interaction with Endogenous Networks." *Political Analysis* 20(3):316–328.

Shulman, Stuart. 2011. DiscoverText: software training to unlock the power of text. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*. College Park: ACM Press.

Sprague, John & Robert Huckfeldt. 1992. "Political Parties and Electoral Mobilization: Political Structure, Social Structure, and the Party Canvass." *American Political Science Review* 86(1):70–86.

Steinert-Threlkeld, Zachary C. 2016. "Replication Data for: Longitudinal Network Centrality Using Incomplete Data.".
URL: *http://dx.doi.org/10.7910/DVN/KKWB4A*

Steinert-Threlkeld, Zachary C. 2017. "Longitudinal Network Analysis with Incomplete Data." *Political Analysis* .

Steinert-Threlkeld, Zachary C., Delia Mocanu, Alessandro Vespignani & James Fowler. 2015. "Online social networks and offline protest." *EPJ Data Science* 4(1):19.

Veenhoven, Ruut. 1991. "Is Happiness Relative?" *Social Indicators Research* 24(1):1–34.

Watts, Duncan J. & Peter Sheridan Dodds. 2007. "Influentials, Networks, and Public Opinion Formation." *Journal of Consumer Research* 34(December):441–458.

Zachary, Wayne W. 1977. "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research* 33(4):452–473.

Zeitzoff, Thomas, John Kelly & Gilad Lotan. 2015. "Using social media to measure foreign policy dynamics: An empirical analysis of the Iranian-Israeli confrontation (2012-13)." *Journal of Peace Research* 52(3):368–383.

Notes

[1]Distance measures the number of steps between two nodes. If A and B are connected to each other, their distance is one. If A is connected to C through B, the distance AC is two while the distance BC is 1.

[2]A k-core is a subset of nodes in the network in which all nodes have at least degree k.

[3] For replication code of Figure 1, see this paper's repository at Harvard Dataverse (Steinert-Threlkeld 2016). That repository contains replication code and data for the rest of the paper as well.

[4]See Section 5 of the Supplementary Materials for a discussion of why Twitter is so popular, limitations of working with is data, and non-follower networks that Twitter data can measure.

[5] "Following" is the fundamental building block of Twitter. When this paper refers to two accounts in a relationship or a connection existing between two accounts, it means that one follows the other.

[6]Twitter is an asymmetric network, and the terms "follower" and "friend" have different meanings. A "follower" is an account which has indicated to Twitter that it wants to automatically be made aware of the tweets of an account it follows, while a "friend" is the account being followed. If B follows A, B is A's follower while A is B's friend.

[7]Twitter provides an account's followers list via the GET followers/ids or GET followers/list REST API endpoints. Twitter's GET users/ids endpoint of its REST API provides users' metadata. This information includes the user's screen name, self-reported location, preferred language, number of friends, number of followers, and, most importantly, date the user joined Twitter. Passing the ID numbers from GET followers/ids to GET users/ids is faster than downloading the followers via GET followers/list.

[8]Because online social networks grow by registering new users, population growth on the platforms academics study is higher than actual population growth.

[9]Section 4 of the Supplementary Materials present a colored version of Figure 4b.

[10] "Group fixed effects" refers to the fact that an account is known to belong to 1 of 4 social movements, each with a different average number of followers at the start of the study.

[11]Coordination is measured using a Bernoulli Naive Bayes topic model. For detail on the construction of that model, see the Supplementary Materials of Steinert-Threlkeld (2017).