amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Diffusion Centrality: Foundations and Extensions

Yann Bramoullé
Garance Genicot

WP 2018 - Nr 37

Diffusion Centrality: Foundations and Extensions

Yann Bramoullé and Garance Genicot*

October 2018

Abstract: We first clarify the precise theoretical foundations behind the notion of diffusion centrality. This allows us to address a minor inconsistency in the model description of Banerjee et al. (2013). We then identify unnatural implicit assumptions in the model of political intermediation proposed by Cruz, Labonne & Querubín (2017). We introduce two extensions of diffusion centrality, *targeting centrality* and *reachability*, which we believe better capture features of contexts with targeted requests. We derive general explicit formulas to compute these new measures.

Keywords: Diffusion centrality, Katz-Bonacich centrality, political intermediation, targeting.

*Bramoullé: Aix-Marseille Univ., CNRS, EHESS, Centrale Marseille, AMSE; Genicot: Department of Economics, Georgetown University. For financial support, Yann Bramoullé thanks the European Research Council (Consolidator Grant n. 616442). Garance Genicot is grateful to the Aix-Marseille School of Economics for their warm hospitality during the writing of this paper.

I Introduction

In a groundbreaking study, Banerjee et al. (2013) propose a new measure, diffusion centrality, to capture the extent to which a piece of information given to an agent eventually diffuses in a network. Diffusion centrality is the expected number of times all agents hear about the information, in a simple model of information diffusion. It admits a simple explicit expression related to a discounted sum of powers of the network's adjacency matrix and nests the well-known eigenvector and Katz-Bonacich centralities. Banerjee et al. (2013) show that diffusion centrality performs well empirically by using it to explain take-up rates of a microfinance loan program in rural India. Two important recent papers build on this notion. Banerjee et al. (2018) show how diffusion centrality can help explain agents' ability to identify others' effectiveness at diffusing information. In a context of political competition, Cruz, Labonne & Querubín (2017) propose a theory of political intermediation where citizens' requests for favors from politicians are transmitted through the social network. Their theory predicts that candidates with higher diffusion centrality should receive more votes. They show empirically that the eigenvector centrality of a politician in the family network is indeed associated with a higher vote share in data from the Philippines.

We identify and address some modelling inconsistencies in these three papers, and obtain three main results. *We first clarify the precise theoretical foundations behind the notion of diffusion centrality.* Contrary to what is claimed in Banerjee et al. (2013) and Banerjee et al. (2018), diffusion centrality does not emerge from a model where "at each iteration every informed node tells each neighbor with probability q", Banerjee et al. (2013, p.1236498-6).[1] Rather, only nodes receiving the information at $t-1$ have any likelihood of transmitting it to their neighbors at t. In other words, informed nodes transmit the information only if they received it in the previous period. Further, we show that diffusion centrality relies on an additional implicit assumption: if an agent receives the information from k different sources in period $t-1$, she must transmit the information independently k times to her neighbors in period t. Relaxing either assumption leads to different computations and

[1] Similarly, Banerjee et al. (2018) write on p.18-19: "In each period, with probability $w_{ij} \in (0,1]$, independently across pairs of neighbors and history, each informed node i informs each of its neighbors j of the piece of information and the identity of its original source."

centrality notions.

We then assess the theoretical approach of Cruz, Labonne & Querubín (2017). We show that their model relies on two important implicit assumptions. When an agent sends a request for a favor to a politician, this politician must provide a new favor every time she hears about the request. Moreover, this targeted request is retransmitted by the same politician and to the same agent during information diffusion. We argue that both assumptions are quite unnatural in this context. In addition, they imply that the number of favors provided is infinite under the parametrization used in the empirical analysis, see Section 3. We propose two changes.

First, we propose to adapt the measure of diffusion centrality to targeted requests for favors. We assume that the request for a favor is not retransmitted by the request's target nor to the request's initiator during information diffusion. We define the targeting centrality of an agent as the expected number of times this agent receives other people's requests for favors under these natural assumptions of no retransmission. *This yields our second result: we derive an explicit formula for targeting centrality.* We show that this new measure has similar computational complexity, but differs significantly from diffusion centrality in some contexts.[2]

Second, we assume that the agent sends a request for a specific favor, which can therefore be granted only once. Her expected utility is then proportional to the probability that her request will reach the politician. Theoretically, these probabilities are not simply related to the expected number of times the politician hears about the request. *In our third result, we provide a general formula to compute the reachability of an agent - the sum of the probabilities that targeted messages will reach her - based on the inclusion - exclusion principle.* This formula is combinatorially complex, which confirms that computing these probabilities is computationally hard. Obtaining numerical approximations of arbitrary precision is straightforward but can still be computationally intensive.

Alternatively and as proposed by Banerjee et al. (2013), researchers could make use of

[2]Our formula is valid under the assumption that the number of periods is infinite, which is a maintained assumption of Cruz, Labonne & Querubín (2017). Our measure thus extends Katz-Bonacich and eigenvector centrality.

proxies: simpler measures which are easier to compute and may be highly correlated with reachability. Proxies' usefulness varies, however, and may depend on specific features of the context. We conjecture that targeting centrality will provide a significantly better proxy than diffusion centrality in situations of targeted requests like political intermediation.

II Foundations

We consider a standard model of information diffusion in a network, as in Banerjee et al. (2013, 2018) and Cruz, Labonne & Querubín (2017). A finite number n of agents are embedded in a fixed social network G, where $g_{ij} = 1$ if i is linked with j and $g_{ij} = 0$ otherwise. Information is transmitted in discrete iterations. In period 0, one agent initially has the relevant piece of information. In period t, agents in a specific subset transmit the information independently to each of their neighbors with probability α and these stochastic transmissions may independently occur several times within the period. Information diffusion ends in period T.

To fully specify the process of information diffusion, we must make assumptions on which agents transmit information and how many times these agents transmit the information within a period. On the first feature, two alternative assumptions are possible.

Assumption (IS): Information as a stock. *All informed agents transmit the information at t.*

Assumption (IF): Information as a flow. *Only agents who received the information at $t-1$ transmit it at t.*

Assumption (IS) says that every informed agent transmits the information at t. Thus, agents who received the information for the first time in $t-1$ behave like agents who received the information in earlier periods. Transmission of information in a given period then depends on the stock of informed agents. By contrast, only agents who just received the information transmit it under assumption (IF). In this case, receiving the information, again or for the first time, prompts its retransmission. Informed agents who do not receive

the information again in $t-1$ do not send it in period t, and transmission of information now depends on the flow of informed agents.

On the second feature, we again distinguish between two natural assumptions.

Assumption (UM): Unique retransmission of multiple signals. *If agent i receives the information from distinct sources at $t-1$, she retransmits the information only once at t.*

Assumption (MM): Multiple retransmission of multiple signals. *If agent i receives the information from S distinct sources at $t-1$, she retransmits the information independently S times at t.*

Next, we introduce some notions and notations. Let $n_{ij}(T)$ denote the expected number of times agent j hears about the information within the first T periods when the information is initially given to agent i. Let $n_{ij} = \lim_{T\to\infty} n_{ij}(T)$. Define $n_i(T) = \sum_j n_{ij}(T)$ as the expected number of times all agents hear about the information when it is initially given to i, and $n_i = \lim_{T\to\infty} n_i(T)$. A walk of length t in G connecting i to j is a set of t agents $i_1 = i, i_2, ..., i_t = j$ such that $\forall s < t, g_{i_s i_{s+1}} = 1$. Let $W_{ij}(T)$ be the set of walks connecting i to j in G and of length less than or equal to T and let W_{ij} be the set of all walks connecting i to j. Given walk w, let $l(w)$ be the length of the walk. Given two walks w and w', let $w \cap_b w'$ denote the intersection of the beginning of the two walks. For instance if $w = \{iklj\}$ and $w' = \{ikmj\}$, $w \cap_b w' = \{ik\}$. Our first result characterizes precisely when $n_i(T)$ is equal to diffusion centrality.

Proposition 1 *Consider a model of information transmission in discrete iterations. Under assumptions (IF) and (MM),*

$$n_{ij}(T) = \sum_{w \in W_{ij}(T)} \alpha^{l(w)} = \sum_{t=1}^{T} \alpha^t [G^t]_{ij}.$$

Under assumptions ((IS) and (MM)) or ((IF) and (UM)), this equality does not generally hold.

Proof: Take any two walks of size t originating in i and ending in j, w and w'. Let e be the node at the end of intersection $w \cap_b w'$ ($e = i$ if the intersection is empty). Under (IF) the probability that the message will reach e is $\alpha^{l(w \cap_b w')}$. Once the walks separate, it follows from (MM) that two independent messages travel among the remaining parts of the two walks. Conditional on reaching e, one message has probability $\alpha^{l(w)-l(w \cap_b w')}$ of reaching j and the other reaches j with probability $\alpha^{l(w')-l(w \cap_b w')}$. Thus, the expected number of messages from i that reach j over these two walks is $\alpha^{l(w \cap_b w')}(\alpha^{l(w)-l(w \cap_b w')} + \alpha^{l(w')-l(w \cap_b w')}) = \alpha^{l(w)} + \alpha^{l(w')}$. Generalizing the argument to any number of walks tells us that $n_{ij}(T) = \sum_{w \in W_{ij}(T)} \alpha^{l(w)}$. Since $[G^t]_{ij}$ is the number of walks of length t originating in i and ending in j, $n_{ij}(T) = \sum_{t=1}^{T} \alpha^t [G^t]_{ij}$.

Under (IS), the expected number of messages is generally higher than provided by diffusion centrality. For instance, consider the line $1 - 2 - 3$ with $T = 2$. Under (IF), we have $n_{11}(2) = \alpha^2$, $n_{12}(2) = \alpha$ and $n_{13}(2) = \alpha^2$, leading to a diffusion centrality of $n_1(2) = \alpha + 2\alpha^2$ for agent 1. By contrast under (IS), agent 1 may retransmit to agent 2 at period 2. This now yields $n_{12}(2) = 2\alpha$ and $n_1(2) = 2\alpha + 2\alpha^2$.

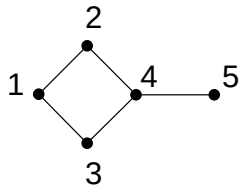

Figure 1. Diffusion centrality can necessitate multiple retransmissions of signals.

Next, consider the network depicted in Figure 1 and $T = 3$. When agent 1 initially sends the information, at time 2 agent 4 may receive messages both from agent 2 and agent 3. Diffusion centrality yields $n_{15}(3) = \alpha^3 [G^3]_{15} = 2\alpha^3$, which assumes that agent 3 retransmits both messages in period 3. By contrast, $n_{15}(3) = \alpha^3$ under (UM). QED.

Proposition 1 clarifies the theoretical foundations behind diffusion centrality. As discussed in the Introduction, this result allows us to correct a minor inconsistency in the descriptions of the models in Banerjee et al. (2013, 2018). The neat interpretation of diffusion centrality only holds in a setup where recently informed nodes retransmit the information. This indicates yet another difference between the models of information transmission

underlying the two centrality notions proposed by Banerjee et al. (2013). Note that information is treated as a stock rather than a flow in the process of information diffusion that underlies their structural estimations and that gives rise to communication centrality.[3] Therefore, communication centrality relies on assumption (IS), while diffusion centrality relies on assumption (IF).

III Extensions: Diffusion of Targeted Requests

A Motivation

We next discuss and assess the theoretical approach developed in Cruz, Labonne & Querubín (2017). The authors build a model of political intermediation, where a request for a favor by a citizen is transmitted to the elected politician through the social network. In the model, two candidates A and B compete for votes. Voter i derives utility U_i^A from the clientelistic goods and services received from candidate A if elected. We reproduce two key paragraphs of the paper.[4]

"We assume that requests for goods and services are passed on through the social network. If voter i wants to receive a clientelistic good, she needs to enlist the help of intermediaries that will connect her personally to the incumbent. Let α be the probability that each intermediary passes on the request successfully. A walk of length m between voter i and candidate A will yield the desired outcome (i.e., the favor, good, or service will be provided) with probability α^m. The voter derives utility b from accessing the service. Thus, voter i's expected access to clientelistic goods and services is a decreasing function of the network distance between her and the elected candidate.

The social network is captured by the adjacency matrix G. The elements (g_{ij}) of the matrix take a value one if i and j are connected and zero otherwise. The elements of G^m,

[3] "In each subsequent period, households that have been informed in previous periods pass information to each of their neighbors, independently, with probability q_P if they are participants and with probability q_N if they are not.", Banerjee et al. (2013, 1236498-2).

[4] The authors adopt notation A to denote both a candidate and the adjacency matrix of the network. For clarity and consistency, we denote the adjacency matrix by G, its elements by g_{ij} and the number of walks of length m by $g_{ij,m}$ and make use of these notations when reproducing the paragraphs.

denoted $(g_{ij,m})$, capture walks of length m between i and j. Taking all potential walks into account, voter i's utility if A is elected is given by

$$U_i^A = b \sum_{m=1}^{\infty} g_{iA,m} \alpha^m.$$

Vote share becomes:

$$\begin{aligned} VS^A &= \frac{1}{2} + \frac{b}{2N\sigma} \sum_i (\sum_{m=1}^{\infty} g_{iA,m}\alpha^m - \sum_{m=1}^{\infty} g_{iB,m}\alpha^m) \\ VS^A &= \frac{1}{2} + \frac{b}{2N\sigma} \sum_{m=1}^{\infty} (\sum_i g_{iA,m}\alpha^m - \sum_i g_{iB,m}\alpha^m).\text{", Cruz, Labonne \& Querubín (2017, p. 3010).} \end{aligned}$$

These computations and Proposition 1 imply that A's vote share is an affine function of $\sum_i n_{iA}$ defined above (under assumptions (IF) and (MM)). Since G is symmetric, $n_{ij}(T) = n_{ji}(T)$ and $\sum_i n_{iA} = \sum_i n_{Ai} = n_A$ and hence vote share simply depends on the candidates' Katz-Bonacich centrality: $VS^A = \frac{1}{2} + \frac{b}{2N\sigma}(n_A - n_B)$. This is the central theoretical prediction that the authors bring to data.

We observe several inconsistencies in this description. Note that a voter's benefit is linear in the number of requests successfully passed to the incumbent. These computations therefore implicitly assume that every successful request translates into a new favor. This is inconsistent with voter i asking for a specific favor. In addition, Cruz, Labonne & Querubín (2017) further assume that $\alpha = 1/\lambda_{\max}(G)$ where $\lambda_{\max}(G)$ is G's largest eigenvalue. In that case, the number of successful requests diverges to infinity and this model predicts that politicians will provide an infinite number of favors.[5]

Another, perhaps more subtle, issue concerns the way information diffuses in the network. In this context of political intermediation, note that the piece of information transmitted by agents explicitly mentions the identities of the request's sender and target. It corresponds to the statement: "Agent i needs a favor from politician A". Current com-

[5]Cruz, Labonne & Querubín (2017) incorrectly claim on p.3011: "For this particular value of α, Katz centrality is equal to eigenvector centrality". As $\alpha \to 1/\lambda_{\max}(G)$, Katz-Bonacich centrality n_A diverges to infinity. It is the ratio n_A/n_B which converges to the ratio of the eigenvector centralities of the two candidates.

putations assume that this information is retransmitted by A and to i during information diffusion. However, it seems more sensible to assume that the request is not retransmitted by its target nor to its initiator during diffusion.

The approach proposed by Cruz, Labonne & Querubín (2017) has great merit. The insight that political competition depends on the intermediation of favors in the network seems important, and deserves proper theoretical foundations. We thus propose to modify their framework as follows. In a first stage, we assume that a request for a favor is not retransmitted to its sender nor by its target during information diffusion. We obtain an explicit formula for the expected number of times a politician hears about a citizen's request. This leads to a new centrality measure, which entails similar computational complexity but differs significantly from diffusion centrality.

In a second stage, we relax the assumption that every successful request leads to a new favor. Rather, we assume that an agent sends a request for a specific favor, which can be granted only once. An agent's expected utility is then proportional to the probability that the incumbent will hear about her request.[6] We provide a general formula to compute these probabilities, based on the inclusion - exclusion principle.

B Targeting Centrality

Consider a model of information transmission, as in Section 2 ,where i sends her request for a favor in period 1 and T tends to infinity. From Proposition 1, we know that under (IF) and (MM) $n_{iA} = \sum_{w \in W_{iA}} \alpha^{l(w)}$ counts the expected number of times the request reaches the incumbent and Cruz, Labonne & Querubín (2017) assume that U_{iA} is proportional to n_{iA}. Next, assume that if an agent other than A receives the request in period t, she transmits it independently to each of her neighbors except for i in period $t+1$, and if A receives the request in period t, she does not retransmit it. Thus, the request is not retransmitted by its target A nor to its sender i. Denote by $\bar{n}_{iA}$ the expected number of

[6] Alternatively, we could assume that the probability p_{iA} of the request being granted is increasing in the number of times the incumbent hears about it: $p_{iA} = p(n_{iA})$ with $p : \mathbb{R}_+ \to [0,1]$. Under this assumption, vote share depends on $\sum_i p(n_{iA})$, which differs, again, from diffusion centrality and now also depends on properties of the function $p(.)$.

times i's request reaches A under these alternative assumptions. Let $\bar{W}_{iA}$ denote the set of walks connecting i to A and such that i only appears at the beginning of the walk and A only appears at the end of the walk. Denote by $G[i]$ the network over $n-1$ nodes obtained by removing i and her links.

Proposition 2 *Suppose that i's request for a favor from $A \neq i$ diffuses in the network under assumptions (IF) and (MM) and that i's request is not retransmitted to i or by A during information diffusion. The expected number of times i's request reaches A is equal to*

$$\bar{n}_{iA} = \sum_{w \in \bar{W}_{iA}} \alpha^{l(w)} = \frac{n_{iA}(G)}{[1 + n_{ii}(G)][1 + n_{AA}(G[i]))]} = \frac{n_{iA}(G)}{[1 + n_{ii}(G[A])][1 + n_{AA}(G)]}.$$

Proof: A direct application of the arguments of Proposition 1's proof shows that $\bar{n}_{iA} = \sum_{w \in \bar{W}_{iA}} \alpha^{l(w)}$. Next, a cycle originating at i is a walk from i to i. Let $C_i(G)$ denote the union of the set of cycles originating at i in network G and of the empty cycle, and similarly for $C_A(G)$. Each walk from i to A can be uniquely decomposed into: a cycle from i to i (possibly empty), a walk where i only appears at the beginning and A only appears at the end, and a cycle from A to A which does not go through i (possibly empty). This decomposition implies that:

$$\begin{aligned} n_{iA} &= \sum_{w \in W_{iA}(G)} \alpha^{l(w)} = \sum_{c \in C_i(G), w' \in \bar{W}_{iA}(G), c' \in C_A(G[i])} \alpha^{l(c)+l(w')+l(c')} \\ &= \Big(\sum_{c \in C_i(G)} \alpha^{l(c)}\Big)\Big(\sum_{w' \in \bar{W}_{iA}(G)} \alpha^{l(w')}\Big)\Big(\sum_{c' \in C_A(G[i])} \alpha^{l(c')}\Big). \end{aligned}$$

Since $C_i = W_{ii} \cup \{\varnothing\}$, $\sum_{c \in C_i(G)} \alpha^{l(c)} = 1 + \sum_{w \in W_{ii}(G)} \alpha^{l(w)} = 1 + n_{ii}(G)$ and similarly $\sum_{c' \in C_A(G[i])} \alpha^{l(c')} = 1 + n_{AA}(G[i])$.

Similarly, each walk from i to A in G can be uniquely decomposed into: a cycle from i to i which does not go through A (possibly empty), a walk where i only appears at the beginning and A only appears at the end, and a cycle from A to A (possibly empty), leading to the second equality. QED.

Proposition 2 shows that the expected number of times the request reaches the incumbent when the request is not retransmitted by A or to i is simply related to this number in the absence of constraints on retransmission. The discounted number of walks connecting i to A when i appears only at the beginning and A appears only at the end is equal to the discounted number of unconstrained walks connecting i to A divided by the discounted number of cycles starting at i and by the discounted number of cycles starting at A in $G[i]$. Thus, if $A \neq i$, [7]

$$\bar{n}_{iA} = \frac{[\alpha G(I - \alpha G)^{-1}]_{iA}}{[I - \alpha G]_{ii}^{-1}[I - \alpha G[i]]_{AA}^{-1}}$$

By definition, the *targeting centrality* of A is $\bar{n}_A = \sum_i \bar{n}_{iA}$ and is equal to the expected number of times A hears about any citizen's request under these no-retransmission assumptions. Targeting centrality differs from diffusion centrality $n_A = \sum_i n_{iA}$. Since it also relies on elements on inverse matrices of the kind $(I - \alpha M)^{-1}$, however, its computational complexity is of the same order of magnitude as for diffusion centrality.

We explore the relation between diffusion centrality and this new measure through numerical simulations. We consider Erdös-Renyi random graphs with $n = 50$ agents and probability of link formation $p = 0.2$. We pick $1,000$ graphs at random and for each graph, we compute how the correlation between the two measures varies with α. For small values of α, only direct links carry weight and targeting and diffusion centrality will hardly differ. The question is: what happens as α increases? We depict the results in Figure 2. Note that diffusion centrality is only well-defined for $\alpha < \alpha_{\max} = 1/\lambda_{\max}(G)$ and we represent the ratio $\alpha/\alpha_{\max}$ on the x axis. We represent how the 5^{th}, 50^{th} and 95^{th} percentiles of the distribution of correlations vary with $\alpha/\alpha_{\max}$.

[7] A direct implication of Proposition 2 is that for any i, j,

$$[I - \alpha G]_{ii}^{-1}[I - \alpha G[i]]_{jj}^{-1} = [I - \alpha G]_{jj}^{-1}[I - \alpha G[j]]_{ii}^{-1}.$$

To our knowledge, this provides a novel result in matrix analysis.

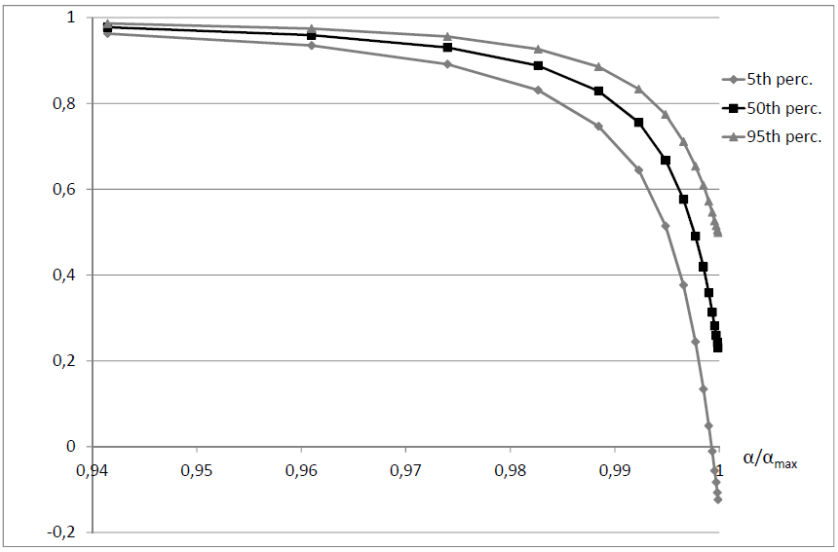

Figure 2. Correlation between targeting and diffusion centrality.

We see that n_A and $\bar{n}_A$ are initially very highly correlated but that this close relation breaks down when α gets close to its maximal value. Correlation is generally decreasing and concave in α. For values very close to $1/\lambda_{\max}(G)$, correlation between the two measures tends to be quite low, about 0.23 for the median network, and displays significant dispersion, from -0.12 at the 5^{th} percentile to 0.50 at the 95^{th} percentile.[8] The correlation between the two measures is therefore strongly affected by the structure of the network, and can even be negative.

Which measure provides a better proxy is, ultimately, an empirical question and hence likely depends on the specific context studied. We conjecture that targeting centrality will perform better in contexts where the transmitted information explicitly mentions the identities of sender and target and, in particular, in the context of political competition studied in Cruz, Labonne & Querubín (2017).

In other contexts, the information transmitted may mention the identity of the sender only or of the target only. In these cases, we can adapt our arguments above to obtain corresponding corrections. For instance, in the gossip model of Banerjee et al. (2018), the transmitted information is of the following kind: "Agent i first said that [...]". Under the assumption that this information is not retransmitted to i, the expected number of times

[8] Banerjee et al. (2013, 2018) assume that $\alpha = 1/\lambda_{\max}(G)$ in their empirical implementation of diffusion centrality with $T < \infty$. Cruz, Labonne & Querubín (2017) also assume that $\alpha = 1/\lambda_{\max}(G)$.

i's gossip reaches j is equal to $n_{ij}/[1+n_{ii}] = [\alpha G(I-\alpha G)^{-1}]_{ij}/[I-\alpha G]_{ii}^{-1}$ rather than n_{ij}. Similarly, if the information is not retransmitted by j, the expected number of times i's information reaches j is equal to $n_{ij}/[1+n_{jj}] = [\alpha G(I-\alpha G)^{-1}]_{ij}/[I-\alpha G]_{jj}^{-1}$.

C Reachability

Finally, assume that an agent sends requests for a specific favor, that can be granted only once. Denote by $p_{iA}(T) \in [0,1]$ the probability of i's favor request successfully reaching A within the first T periods and $p_{iA} = \lim_{T\to\infty} p_{iA}(T)$. The expected utility of agent i is now equal to $U_{iA} = p_{iA}b$. Define the *reachability* of A as $\sum_i p_{iA}$, the overall expected number of favors provided by the incumbent when every agent sends a request. The vote share of a candidate is now a simple affine function of her reachability. The computation of these probabilities depends on details of the underlying process of request transmission. We assume (IF), (MM) and no retransmission by the target or to the sender in what follows; our result below can be extended to alternative assumptions.

Not surprisingly, reachability does not display the additive property underlying the counting formula assumed in Cruz, Labonne & Querubín (2017). For instance, suppose that i has a direct connection with A and an indirect connection through common friend j. In this case, $p_{iA} = \alpha + (1-\alpha)\alpha^2 = \alpha + \alpha^2 - \alpha^3$. We next derive a general formula based on the inclusion - exclusion principle, and which could be used for algorithmic implementation. Let us introduce some notions and notations. Denote by $w_1 \cap_b w_2 \cap_b ... \cap_b w_k$ the intersection of the beginning of the k walks $w_1, w_2, ..., w_k$. Define $L(w_1, ..., w_k)$ as follows:

$$L(w_1, ..., w_k) = \sum_{s_1=1}^{k} l(w_{s_1}) - \sum_{s_1<s_2} l(w_{s_1} \cap_b w_{s_2}) + ... + (-1)^{k+1} l(w_1 \cap_b ... \cap_b w_k).$$

The general idea here is to count common beginnings only once. As soon as two walks separate, however, we add the lengths of the remaining segments. For instance with $k=2$, $L(w_1, w_2) = l(w_1) + l(w_2) - l(w_1 \cap_b w_2)$. Finally, write $w_1 \neq ... \neq w_k$ to denote that the k walks $w_1, ..., w_k$ are distinct. Let $\bar{W}_{iA}(T)$ denote the set of walks connecting i to A such that i only appears at the beginning of the walk and A only appears at the end of the

walk, and of length less than or equal to T.

Proposition 3 *Consider a model of political intermediation under (IF), (MM) and the assumption that i's request for a favor from A is not retransmitted to i or by A. The probability of i's request successfully reaching A is equal to*

$$p_{iA}(T) = \sum_{k=1}^{|\bar{W}_{iA}(T)|} (-1)^{k+1} \sum_{w_1 \neq ... \neq w_k \in \bar{W}_{iA}(T)} \alpha^{L(w_1,...,w_k)}.$$

Proof: We consider requests eventually reaching A from a walk in $\bar{W}_{iA}(T)$ as events. We know that the request reaches its target if and only if it reaches it through such a walk. We can than apply the principle of inclusion - exclusion to these events. This implies that

$$p_{iA}(T) = \sum_{k=1}^{|\bar{W}_{iA}(T)|} (-1)^{k+1} \sum_{w_1 \neq ... \neq w_k \in \bar{W}_{iA}(T)} p(w_1 \wedge ... \wedge w_k)$$

where $p(w_1 \wedge ... \wedge w_k)$ is the probability of the info reaching A through all walks $w_1, w_2, ..., w_k$. Next, let us show recursively that $p(w_1 \wedge ... \wedge w_k) = \alpha^{L(w_1,...,w_k)}$.

For any walk w, the probability that a request will travel all the way through w is $p(w) = \alpha^{l(w)}$. Next, take two walks $w \neq w'$ originating in i. The chain rules tells us that $p(w \wedge w') = p(w)\, p(w'|w)$ where $p(w'|w)$ is the probability that i's request will go through w' conditional on having gone through w.

$$p(w'|w) = p(w' \setminus (w \cap_b w')|w) = \alpha^{l(w')-J(w,w')}$$

with $J(w, w')$ denoting the number of links initially common to w and w': $J(w, w') = l(w \cap_b w')$. Hence,

$$p(w \wedge w') = \alpha^{l(w)+l(w')-l(w \cap_b w')} = \alpha^{L(w,w')}.$$

Assume now that we have proven that $p(w'_1 \wedge ... \wedge w'_{k-1}) = \alpha^{L(w'_1,...,w'_{k-1})}$ for any set of $k-1$ walks originating in i ($k \geq 2$) and that we have constructed $J((w'_1 \wedge ... \wedge w'_{k-2}), w)$ for any $\{w_1, ..., w_{k-2}\}$ and w that originates in i.

Take a set of k walks originating in i: $\{w_1, ..., w_k\}$. Clearly,

$$p(w_1 \wedge ... \wedge w_k) = p(w_1 \wedge ... \wedge w_{k-1})\ p(w_k | w_1 \wedge ... \wedge w_{k-1}). \tag{1}$$

The probability that a request will go through w_k conditional on having gone through all walks in $\{w_1, ...w_{k-1}\}$ is the probability that it will go through the remainder of w_k once we remove the initial links that may have been accounted for:

$$p(w_k | w_1 \wedge ... \wedge w_{k-1}) = \alpha^{l(w_k) - J((w_1 \wedge ... \wedge w_{k-1}), w_k)} \tag{2}$$

where $J((w_1 \wedge ... \wedge w_{k-1}), w_k)$ is the number of links initially common to $((w_1 \wedge ... \wedge w_{k-1}))$ and w_k. Notice that $J((w_1 \wedge ... \wedge w_{k-1}), w_k)$ equals the number of links initially common to both w_k and $(w_1 \wedge ... \wedge w_{k-2})$ plus the number of links initially common to both w_k and w_{k-1} minus the double counting:[9]

$$J((w_1 \wedge ... \wedge w_{k-1}), w_k) = J((w_1 \wedge ... \wedge w_{k-2}), w_k) + J(w_{k-1}, w_k) - J((w_1 \wedge ... \wedge w_{k-2}), w_{k-1} \cap_b w_k)). \tag{3}$$

Using (2) in (1) along with our induction hypothesis tells us that

$$p(w_1 \wedge ... \wedge w_k) = \alpha^{L(w_1, ..., w_{k-1})}\ \alpha^{l(w_k) - J((w_1 \wedge ... \wedge w_{k-1}), w_k)}.$$

Expanding the terms in (3), we get

$$\begin{aligned} l(w_k) - J((w_1 \wedge ... \wedge w_{k-1}), w_k) &= l(w_k) - \sum_{s \in \{1, ..k-1\}} l(w_s \cap_b w_k) + ... + (-1)^{k-1} l(w_1 \cap_b ... \cap_b w_k) \quad (4) \\ &= L(w_1, ..., w_k) - L(w_1, ..., w_{k-1}). \quad (5) \end{aligned}$$

It follows that $p(w_1 \wedge ... \wedge w_k) = \alpha^{L(w_1, ..., w_k)}$. QED

Note that the first term in the formula is equal to $\sum_{w \in \bar{W}_{iA}(T)} \alpha^{l(w)}$ which, by a direct extension of the first part of Proposition 2, is equal to $\bar{n}_{iA}(T)$. Proposition 3 then clarifies

[9] The number of links common to $(w_1 \wedge ... \wedge w_{k-2})$ and the initial intersection between w_k and w_{k-1}.

the difference between the number of times A is expected to hear about i's request and the probability that A will hear about it. This formula is combinatorially complex, which confirms that computing these probabilities in practice is computationally hard. Applied researchers then have two options. One is to rely on numerical simulations to obtain approximate values of these probabilities. This is straightforward but can still be computationally intensive. Simply generate N realizations of information diffusion at random and count the number of times K that i's request for a favor reaches A. Then, K/N converges to p_{iA} as N tends to infinity. A similar numerical procedure underlies the structural estimations in Banerjee et al. (2013).

Alternatively, and as proposed by Banerjee et al. (2013), researchers can rely on simpler proxies which are easier to compute and likely to be highly correlated with these probabilities. Proxies vary in their usefulness, however, and the literature still lacks formal results on why and when we should expect diffusion centrality to perform well empirically. In a context of targeted requests, such as political intermediation, we conjecture that targeting centrality may provide a significantly better proxy than diffusion centrality.

REFERENCES

Banerjee, Abhijit, Chandrasekhar, Arun G., Duflo, Esther and Matthew O. Jackson. 2013. "The Diffusion of Microfinance." *Science* 341(6144): 1236498.

Banerjee, Abhijit, Chandrasekhar, Arun G., Duflo, Esther and Matthew O. Jackson. 2018. "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials." Working Paper version of March 12th 2018.

Cruz, Cesi, Labonne, Julien and Pablo Querubín. 2017. "Politician Family Networks and Electoral Outcomes: Evidence from the Philippines." *American Economic Review* 107(10): 3006-3037.