Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Social Networks 27 (2005) 55–71

SOCIAL NETWORKS

www.elsevier.com/locate/socnet

Centrality and network flow ☆

Stephen P. Borgatti *

Department of Organization Studies, Boston College, Carroll School of Management, Chestnut Hill, MA 02467, USA

Abstract

Centrality measures, or at least popular interpretations of these measures, make implicit assumptions about the manner in which traffic flows through a network. For example, some measures count only geodesic paths, apparently assuming that whatever flows through the network only moves along the shortest possible paths. This paper lays out a typology of network flows based on two dimensions of variation, namely the kinds of trajectories that traffic may follow (geodesics, paths, trails, or walks) and the method of spread (broadcast, serial replication, or transfer). Measures of centrality are then matched to the kinds of flows that they are appropriate for. Simulations are used to examine the relationship between type of flow and the differential importance of nodes with respect to key measurements such as speed of reception of traffic and frequency of receiving traffic. It is shown that the off-the-shelf formulas for centrality measures are fully applicable only for the specific flow processes they are designed for, and that when they are applied to other flow processes they get the "wrong" answer. It is noted that the most commonly used centrality measures are not appropriate for most of the flows we are routinely interested in. A key claim made in this paper is that centrality measures can be regarded as generating expected values for certain kinds of node outcomes (such as speed and frequency of reception) given implicit models of how traffic flows, and that this provides a new and useful way of thinking about centrality.

Keywords: Centrality; Network flow; Typology

☆ This paper was first presented at the Sunbelt International Social Networks Conference 2002, in New Orleans. It was subsequently presented at the Complexity colloquium series at Harvard University, in April 2002, and at the Institute for Mathematics and its Applications (IMA) in Minneapolis, in November 2003.
* Tel.: +1 617 552 0450; fax: +1 978 456 7373.
E-mail address: borgatts@bc.edu.

0378-8733/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2004.11.008

1. Introduction

Centrality is one of the most studied concepts in social network analysis. Numerous measures have been developed, including degree centrality, closeness, betweenness, eigenvector centrality, information centrality, flow betweenness, the rush index, the influence measures of Katz (1953), Hubbell (1965), and Hoede (1978), Taylor's (1969) measure, etc.

What is not often recognized is that the formulas for these different measures make implicit assumptions about the manner in which things flow in a network.[1] For example, some measures, such as Freeman's closeness and betweenness (Freeman, 1979), count only geodesic paths, apparently assuming that whatever flows through the network moves only along the shortest possible paths. Other measures, such as flow betweenness (Freeman et al., 1991), do not assume shortest paths, but do assume proper paths in which no node is visited more than once. Still other measures, such as Bonacich's (1987, 1991) eigenvector centrality and Katz's (1953) influence, count walks, which assume that trajectories can not only be circuitous, but also revisit nodes and lines multiple times along the way. Regardless of trajectory, some measures (e.g., betweenness) assume that what flows from node to node is indivisible (like a package) and must take one path or another, whereas other measures (e.g., eigenvector) assume multiple "paths" simultaneously (like information or infections).

What happens when we apply a measure that assumes a given set of flow characteristics to a flow with different characteristics? One of two things must happen: either we lose the ability to fully interpret the measure (as when we compute the mean of a nominal-scaled variable) or we get poor answers (as when we use linear regression to predict values of a dependent variable when the relationship is actually non-linear).

Thus, the immediate objectives of this paper are as follows. First, to construct a list of commonly encountered flow processes and cross-classify them in terms of a few underlying characteristics relevant to measuring centrality. Second, to match existing centrality measures to appropriate kinds of flow processes based on the flow characteristics they assume. Third, to test these ideas by running simulations of flow processes and comparing the results with selected measures of centrality. Fourth, to discuss the development of appropriate methods for flows that currently are not supported by any measures. Fifth, and finally, to discuss how considering flow processes entails a new way to think about centrality that casts centrality measures as expected values in an implicit model of nodal participation in network flows.

2. Typology of flow processes

In this section, I consider a number of commonly encountered flow processes. It is not necessary to enumerate every possible kind of flow, but it is important to generate at least a handful of different kinds. These can then be compared and contrasted in order to elicit dimensions along which they differ. Finally, the dimensions are used to categorize these and

[1] To be more precise, the canonical interpretations we give to these measures are valid to the extent that things flow in certain ways.

other flows. I consider each of the following different kinds of traffic: used goods, money, gossip, e-mail, attitudes, infection, and packages.

2.1. *Used goods*

Consider the case of a used paperback novel that passes from person to person, particularly through the mails. The novel is a solid, indivisible object that can only be in one place at a time. As it goes from person A to person B to person C, etc., it could easily return to a person earlier in the chain, simply because person G has no idea that person B had previously received it, and person B then graciously passes it on to someone else. However, except in special cases (e.g., Alzheimer's), the book does not pass via the same link more than once. That is, if B has sent it to C, and later B receives the book again, he or she will not normally send it to C again.

Hence, the paperback traverses the network using what graph theorist would call a trail—a sequence of incident links in which no link is repeated. Trails are distinguished from paths – sequences in which not only links but also nodes cannot be repeated – and walks, which are unrestricted sequences. All paths are trails and all trails are walks, but not every walk is a trail and not every trail is a path.

2.2. *Money*

Consider a specific dollar bill that moves through the economy, changing hands with each economic transaction. Like the gift, the dollar bill is indivisible and can only be in one place at a time. However, unlike the gift, the dollar bill is not proscribed from passing over the same link more than once. In fact, it could easily move from A to B, B back to A, A to B again, then B to C, and so on. From a graph-theoretic point of view, the bill traverses the network via walks rather than trails. As a result, the movement of money can be modeled as a Markov process.

2.3. *Gossip*

Imagine a juicy, very private, story moving through the informal network of employees of an organization. The story is confidential, which does not impede its flow, but means it is typically told behind closed doors to just one person at a time. Unlike gifts and dollar bills, the story can be in several places at once.[2] It spreads by replication rather than transference. Like gifts but unlike dollar bills, it normally does not pass the same link twice (i.e., I do not tell the same person the same story), but can pass the same node multiple times. Thus, it traces trails through the network rather than walks.

2.4. *E-mail*

A typical example is an e-mail message that warns of an electronic virus. The message is forwarded from one person to several of his contacts, often by sending one message to all

[2] I leave aside the question of whether it is really the "same" object that is in multiple places at once.

of them simultaneously (unlike confidential gossip). The message exists in multiple places at the same time, thanks to diffusion by replication.

2.5. Attitudes

Here, the notion is of an influence process in which, through interaction, individuals effect changes in each other's beliefs or attitudes. Thus, attitudes about what fashion items are "in" versus passé are spread from person to person. The attitudes spread through replication rather than transfer (I do not lose my attitude the moment I infect you with it). A speaker may persuade many people at the same time, and the trajectories followed by the attitude can revisit links—I can continue to influence you about the same thing over time.

2.6. Infection

Consider the case of an infection to which the host becomes immune. The infection spreads from person to person by duplication, like gossip, but does not re-infect anyone who already has had it because they become immune.

2.7. Packages

A package delivery process has the unique characteristic of having a fixed destination or target. In addition, a driver delivering a package normally knows and selects the shortest route possible, so that the package's trajectory follows geodesic paths through a network of roads and intersections.

2.8. Classification

Given these thumbnail sketches, it is not difficult to see a small set of attributes or dimensions along which these different flow processes vary. One attribute has to do with the mechanics of dyadic diffusion: specifically, whether diffusion occurs via replication (copy mechanism) or transfer (move mechanism). Another attribute, applicable only to replication-based flows, is whether the duplication is one at a time (serial), like the passing-on of a paperback novel, or simultaneous (parallel), like a radio broadcast. A third attribute concerns whether the traffic flows non-deterministically, meaning that at any particular juncture, traffic always takes the "best" way (such as taking the shortest possible road to a pre-determined target), or whether traffic flows in a blind, undirected way. Finally, there is an attribute that describes whether trajectories follow graph-theoretic paths, trails, or walks.

The first two attributes both relate to the mechanism of node-to-node transmission. In addition, the second attribute is not independent of the first, since it is only defined for cases falling into one class of the first attribute. As a result, we can simplify the situation by combining the two attributes into a single categorical dimension with three classes: parallel duplication, serial duplication, and transfer.

Similarly, the remaining two attributes are both concerned with the kinds of trajectories that something flowing through the network can take. For convenience, they too can be

Table 1
Typology of flow processes

	Parallel duplication	Serial duplication	Transfer
Geodesics	<No process>	Mitotic reproduction	Package delivery
Paths	Internet name-server	Viral infection	Mooch
Trails	E-mail broadcast	Gossip	Used goods
Walks	Attitude influencing	Emotional support	Money exchange

collapsed into a single categorical dimension that describes the four kinds of trajectories that are realizable. These are geodesics, paths, trails, and walks.

Taken together, these two dimensions can be used to construct a simple typology, as shown in Table 1. In the table, the rows correspond to the trajectory dimension, while the columns correspond to the transmission dimension. The cells of the table correspond to specific flow processes that have been cross-classified by these two dimensions.

Of course, there are other dimensions we can also differentiate among flow processes. For example, some processes might always involve a two-way effect in which A influences B at the same time that B influences A. However, to avoid unnecessary complication, I leave such attributes for exploration at another time.

3. Relation to centrality measures

The purpose of this section is to review a few well-known measures of centrality in order to see what kinds of flow assumptions they make. I begin with closeness centrality.

As defined by Freeman (1979), a node's closeness centrality is the sum of graph-theoretic distances from all other nodes, where the distance from a node to another is defined as the length (in links) of the shortest path from one to the other. In a flow context, we ordinarily interpret closeness as an index of the expected time until arrival of something flowing through the network (Borgatti, 1995). Nodes with low raw closeness scores have short distances from others, and so will tend to receive flows sooner, assuming that what flows originates from all other nodes with equal probability, and also assuming that whatever is flowing manages to travel along shortest paths. In the case of information flows, we normally think of nodes with low closeness scores as being well-positioned to obtain novel information early, when it has the most value. Thus, organizations with low closeness in an R&D technology-sharing network are able to develop products sooner than others. In contrast, individuals with low closeness scores in a sexual network are positioned to catch infections early, possibly before treatments are available in the case of new diseases.

If traffic did not travel along shortest paths, we would not want to interpret closeness as an index of expected time until arrival. Thus, the canonical interpretation of closeness is accurate for two kinds of processes: those in which things flow along shortest paths, such as the package delivery process, and those in which things flow by parallel duplication. In the latter case, all possible paths are followed simultaneously, including the shortest path, and so the net effect is the same. It would be inappropriate to see closeness centrality as an index of reception speed for other flow processes. For example, we might be tempted to use closeness to indicate who is likely to receive news early in a gossip process. However, since gossip does not necessarily follow shortest paths, the rank ordering of who receives information earliest on average will not correspond to the ordering provided by the closeness centrality measure, as can easily be confirmed by simulation. (Tests of this kind are presented in the next section.)

It should also be noted that the shortest path assumption includes a pair of assumptions about reachability. First, the measure only works on connected graphs, since the distance between unconnected nodes is undefined or, popularly, infinite. Second, taking shortest paths implies taking paths that in fact reach a particular destination—what we might call valid paths. If the requirement of taking shortest paths were removed so that traffic could follow any legal graph-theoretic path, we would still need to assume selection of valid paths that actually led from origin to target. The reason for this assumption is that traffic flowing along graph-theoretic paths can easily get stuck in a cul-de-sac from which it could not escape (since paths are defined as a sequence of adjacent nodes in which no node is visited more than once) and never actually reach the target. As a result, in interpreting a closeness measure in terms of time-until-arrival, we implicitly assume a flow process in which traffic from any origin “knows” how to reach any target, much like a non-deterministic computer algorithm.

Another well-known centrality measure is betweenness (Freeman, 1979). Betweenness centrality is defined as the share of times that a node i needs a node k (whose centrality is being measured) in order to reach a node j via the shortest path. Specifically, if g_{ij} is the number of geodesic paths from i to j, and g_{ikj} is the number of these geodesics that pass through node k, then the betweenness centrality of node k is given by

$$\sum_i \sum_j \frac{g_{ikj}}{g_{ij}}, \quad i \neq j \neq k$$

Stated in plain language, betweenness basically counts the number of geodesic paths that pass through a node k. At least, that is the numerator of the measure. The denominator exists to handle the case where there are multiple geodesics between i and j, and node k is only along some of them. Hence, betweenness is essentially k’s share of all paths between pairs that utilize node k—the exclusivity of k’s position. The idea, as Freeman describes it, is that a message traveling from node A to node D in Fig. 1, when confronted with the possibility of taking either route, essentially flips a coin and can be expected to choose the path through B 50% of the time. Thus, betweenness is conventionally thought to measure the volume of traffic moving from each node to every other node that would pass through a given node (Borgatti, 1995). Thus, it measures the amount of network flow that a given node ‘controls’ in the sense of being able to shut it down if necessary.

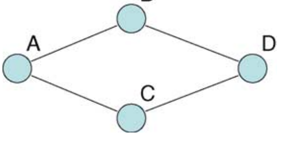

Fig. 1. Traffic flowing from A to D is expected to pass through B or C with equal probability.

What are the assumptions built into this measure? Or, to put it another way, for what kinds of network flows would this measure make sense for? First, it is clear that the measure assumes that the traffic is indivisible. When confronted with several equally short paths, it chooses exactly one at random and proceeds. Thus, the traffic seems to literally move or transfer from node to node, rather than being copied or broadcast from a node. Second, the traffic travels only along shortest paths. Rather than diffusing randomly, it has a target and knows the best way(s) to get there. In fact, by taking all pairs of nodes, the measure systematically takes into account traffic moving from all possible origins to all possible targets.

What kind of flow has these properties? Surely not an infection nor information, which diffuse by copying rather than moving and which do not have targets and do not prefer to take the shortest paths to any node. Thus, it would seem completely inappropriate to use Freeman's centrality measure as an index of the importance of a given node for the spread of infections or the movement of information. Nor does the spread of gossip seem to have these properties. Like infections, gossip is copied rather than moved, and does not normally have a target. Even if it has a target, it is unlikely to reach that target by the shortest possible path, since that would require each node to know what the best route was. Hence, Freeman's centrality measure is probably not ideally suited for measuring a node's ability to control flows of gossip.[3] Rather, the assumptions built in to this measure match the characteristics of the package delivery process, which is characterized by indivisible traffic that transfers from node to node along shortest paths until it reaches a pre-determined target.

Another popular measure of centrality is eigenvector centrality (Bonacich, 1972). Eigenvector centrality is defined as the principal eigenvector of the adjacency matrix defining the network. The defining equation of an eigenvector is

$$\lambda \boldsymbol{v} = \boldsymbol{A}\boldsymbol{v}$$

where $\boldsymbol{A}$ is the adjacency matrix of the graph, λ is a constant (the eigenvalue), and $\boldsymbol{v}$ is the eigenvector. The equation lends itself to the interpretation that a node that has a high eigenvector score is one that is adjacent to nodes that are themselves high scorers. Mathematically, eigenvector centrality is closely related to the measures proposed by Katz (1953), Hubbell (1965), Taylor (1969), Hoede (1978), Coleman et al. (1966), and Friedkin (1991), almost all of which are known as influence measures. The idea is that even if a node influences just one other node, who subsequently influences many other nodes (who themselves influence still more others), then the first node in that chain is highly influential. At the same time, we can see eigenvector centrality as providing a model of nodal risk such that a node's long-term equilibrium risk of receiving traffic is a function of the risk level of its contacts. Hence, a person A in a sexual network may have sex with just one person, but if that person is having sex with many others, the risk to A remains high.

It can be shown (Bonacich, 1987, 1991) that an eigenvector is proportional to the row sums of a matrix $\boldsymbol{S}$ formed by summing all powers of the adjacency matrix, weighted by corresponding powers of the reciprocal of the eigenvalue, as shown in the next equation.

$$\boldsymbol{S} = \boldsymbol{A} + \lambda^{-1}\boldsymbol{A}^2 + \lambda^{-2}\boldsymbol{A}^3 + \dots$$

[3] Except, of course, for nodes that are cut-points, in which case Freeman betweenness provides an exact accounting of node control.

It is also well known that the cells of the matrix powers give the number of walks of length k from node i to node j. Thus, the measure counts the number of walks of all lengths, weighted inversely by length, which emanate from a node.

As a result, the measure “assumes” that traffic is able to move via unrestricted walks rather than being constrained by trails, paths, or geodesics. In addition, the measure does not in any way assume that things flow by transferring or by replicating to one neighbor at a time. Rather, it is consistent with a mechanism in which each node affects all of its neighbors simultaneously, as in a parallel duplication process (first column of the typology). Hence, the eigenvector centrality measure is ideally suited for influence type processes.

Finally, there is degree centrality (Freeman, 1979). Degree centrality can be defined as the number of ties incident upon a node. That is, it is the sum of each row in the adjacency matrix representing the network. To put the measure on the same footing as the other centrality measures discussed here, we can also define degree as the number of paths of length one that emanate from a node. As a result, one way to interpret the measure would be in terms of an implicit process that involves no indirect links. Examples of such processes might be situated knowledge construction in the sense of Lave and Wenger (1991) such that nodes i and j co-construct something that is unique to them—if they engage in similar activity with others, the result would be different and unique to that pair. This describes a kind of flow process that falls outside of the typology presented earlier.

However, another way to interpret the measure is as a measure of immediate effects only—of what happens at time $t+1$ only. For example, if a certain proportion of the nodes in the network are infected with something, and having a tie with an infected node implies getting infected, then the probability of immediate infection is a function of the number of nodes the node is adjacent to. In this sense, degree centrality is similar to eigenvector centrality, the difference being that eigenvector centrality measures a long-term direct and indirect risk while degree centrality measures immediate risk only. By analogy, we can therefore regard degree centrality as a measure of immediate influence—the ability to infect others directly or in one time period. Seen in terms of immediate effects, we can regard degree centrality as appropriate for all parallel duplication flow processes, since in those cases the probability of receiving – in the next time period – something that is randomly distributed in the network, will be entirely a function of the number of ties that a give node has.

Finally, degree models the frequency of visits by something taking an infinitely long random walk through a network (i.e., a money exchange process). As noted earlier, the money exchange process can be modeled as a Markov process. A well-known result in Markov theory is that for a random walk on a graph, the limiting probabilities for the nodes are proportional to degree.[4] Hence, the proportion of times that a node is visited is a function of its degree. This means that degree is an appropriate measure for walk-based transfer processes such as the money exchange process.

[4] This can be shown as follows. Let d_i be the degree of node i, and let $a_{ij}=1$ if i is adjacent to j and 0, otherwise. We define the transition probabilities as $p_{ij}=a_{ij}/d_i$ and observed that $d_j=\sum_i a_{ij}$. If the limiting probabilities, π, are based on degree, then $\pi_i = d_i/\sum d_i$. We then need to show that $\pi P=\pi$. Writing πP as $\sum_i \pi_i p_{ij}$, we can then substitute $d_i/\sum d_i$ for π_i to get $\sum_i [d_i p_{ij}]/\sum d_i$. Recognizing that $d_i p_{ij}$ is a_{ij}, we get $\sum_i a_{ij}/\sum d_i$ which is $d_j/\sum d_i$ which is π_j, the result we were looking for.

Table 2
Flow processes and major centrality measures

	Parallel duplication	Serial duplication	Transfer
Geodesics		Freeman closeness	Freeman closeness Freeman betweenness
Paths	Freeman closeness Freeman degree		
Trails	Freeman closeness Freeman degree		
Walks	Freeman closeness Freeman degree Bonacich eigenvector		

Table 2 locates the four best known measures of centrality and influence in the boxes of the typology for which they are appropriate. It is striking that most of the sociologically interesting processes are not covered by the major measures. The situation improves a little if we consider less well-known and/or forthcoming measures. For example, the influence and status measures of Katz (1953), Hubbell (1965), Taylor (1969), and Hoede (1978) all fall in the [walks, parallel] cell of the table, along with the closeness measure of Stephenson and Zelen (1989). Similarly, Friedkin's (1991) measures of influence also fall in that cell. In contrast, the random walk betweenness and closeness measures of Newman (in press) and Noh and Rieger (2004), respectively, fall into the Markovian [walks, transfer] cell of the table. Still, there are no measures appropriate for infection and gossip processes, which I would regard as extremely important.

4. Simulations

Although not presented formally as a theory, the ideas discussed above do constitute a theory of the structural importance of nodes as a function of flow characteristics. Furthermore, the theory is testable. For example, if I am correct in my assessment of the flow assumptions that underlie betweenness centrality, it should be that when we actually observe traffic moving through a network, the number of times that something passes through a given node should be approximated by the values calculated by the betweenness centrality formula, as long as the traffic moves according to the rules outlined for a package delivery process. For traffic moving in accordance with different rules, such as those of a gossip process, the number of times that traffic moves through a given node should not be well approximated by the formula. In effect, we should be able to see how well the formula does when we apply it to the wrong flow process.

Thus, a critical concept in this section is that of expected versus realized centrality. Expected centrality is the centrality score calculated by the formula that defines it. Realized centrality is the observed value for a node in the context of a particular flow process. In other words, I essentially treat centrality measures as testable models of node importance. For example, in the case of betweenness, the underlying concept being modeled is the

```
For k = 1 to 1000 {independent trials}
        For i = 1 to n {source nodes}
                For j = 1 to n {target nodes} if i <> j then
                        Simulate flow from i to j
        Compute node statistics for this trial
Average the node statistics across trials & compare with centrality measures
```

Fig. 2. Source/target method of running simulations.

amount of traffic that flows through a node. Hence, expected betweenness is a formula-based prediction, and realized betweenness is the actual frequency of traffic we observe flowing through a node across multiple instances. Similarly, the key concept in closeness centrality is the length of time it takes traffic to reach a node (or traffic from a node to reach others). Hence, expected closeness is a formula-based estimate based on path-lengths, and realized closeness is what we obtain when we observe actual flows.

In order to guarantee that the flows strictly follow the rules outlined by the theory, I use simulation to construct observable instances of flow. In general, the simulations proceed as follows. At time 0, I select a source node and assign it a string representing the traffic that will flow, whether by replication or transfer, through the network. At time 1 (and every subsequent time interval), the string flows from any node that has it to one or more (in the case of parallel duplication) of the node's alters, either by copying or moving.

A key question that arises is how to stop the simulation. For the package delivery process (and any other geodesic-based process), there is always a target node, and so the simulation stops when the target has been reached. Thus, a separate simulation is run for each possible combination source and target nodes. In other words, the simulation study has the algorithmic structure as given in Fig. 2.

For all other flow processes, it makes sense to let the simulation run until the string stops moving—i.e., it has reached every node it can reach. This approach would definitely be more in keeping with the notion of modeling real flows. In addition, it is practical in the sense that every path or trail-based process will definitely end somewhere. Only for walk-based measures do we need to create some kind of additional stopping rule (e.g., based on length of walk, or having reached all nodes at least once, etc.), since walks can be infinite. Thus, a separate simulation would be run for each possible source node. This yields the structure as given in Fig. 3.

However, for the purpose of maintaining comparability when validating against Freeman's (1979) centrality measures, I use the source/target method for all flow processes, including ones that do not demand a target. To do otherwise would be to guarantee a lack of fit between the formula-based expected values and the results of the simulation experiments, and could be seen as an unfair comparison. However, it should be clear that in

```
For k = 1 to 1000 {independent trials}
        For i = 1 to n {nodes}
                Simulate flow from i until done moving
        Compute node statistics for this trial
Average the node statistics & compare with centrality measures
```

Fig. 3. Source-only method of running simulations.

order to properly model what happens in most flow processes, a centrality measure should not assume a target node. This topic is taken up in more detail in Section 5.

I do the same with closeness centrality because flow processes without targets are not guaranteed to reach any particular node before becoming trapped. This then means that sometimes the time until arrival is infinite and cannot be averaged in with the rest. In effect, for most real flow processes, the graph traced out by actual flows need not be connected.

Also, it should be noted that in some flow processes, it is not guaranteed that a given simulation from a particular source will actually reach the target. In those cases, the program simply tries again until the target is reached.

Padgett and Ansell's (1993) data on marriages among Renaissance families in Florence are used to illustrate the simulations. The network is shown in Fig. 4. It should be noted that one family was removed because it was an isolate. For simplicity of exposition, only two measures of centrality are discussed: betweenness and closeness.

4.1. *Betweenness centrality*

The first simulation to consider is of a *package delivery* process, whose characteristics are geodesic trajectories achieved by transferring from node to node—or "geodesic + transfer" for short. To simulate the package delivery process, I pick a starting node S and a target node T, and map out all shortest paths from S to T. Then, at time 1, the package moves from S to an adjacent node, chosen randomly with uniform probability, that is along one of the shortest paths. If the package passes through any node F before reaching T, an observed betweenness counter for F is incremented. For each run of the simulation, all combinations

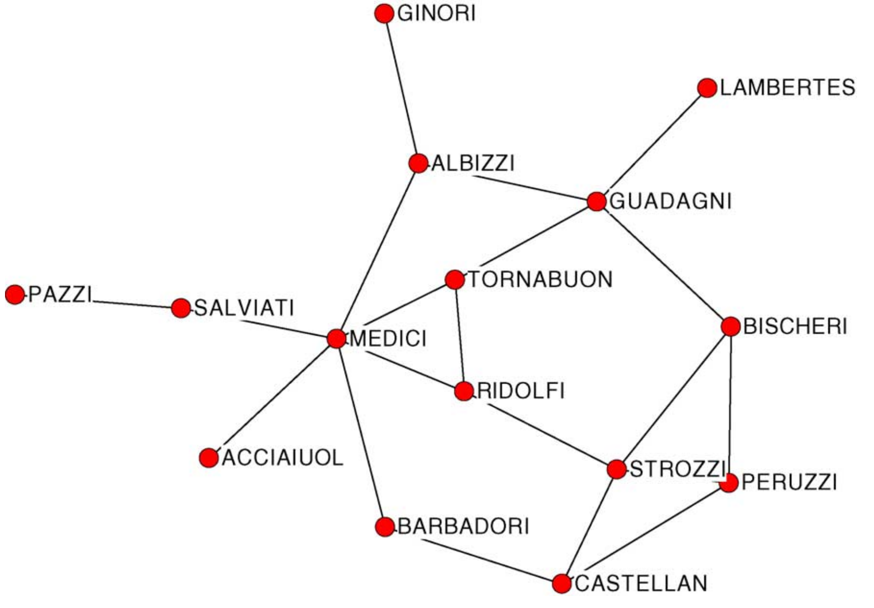

Fig. 4. Padgett's data on marriage ties among Renaissance Florentine families (isolate removed).

of F, S, and T are systematically tried. Finally, each run is repeated 1000 times, and the betweenness tally averaged across runs.

If the theory is correct, then the realized betweenness from the simulation should match exactly the expected betweenness calculated by Freeman's formula. Table 3 (column, "Package") compares the results of 1000 runs of the simulation against the expected value. It is clear that the formula hits the nail on the head, indicating that the flow characteristics of the package delivery process are appropriate for the measure. The question then becomes whether other flow processes are inconsistent with the measure, as predicted above.

The next simulation to consider is the *mooch* process (paths + transfer), in which indivisible traffic is restricted to traveling along paths but is not required to take shortest paths. As noted before, in order to maximize comparability with the betweenness measure, we choose to retain the concept of sources and targets, so that in each run each node serves once as source and as target for each source. For example, when node 1 is the source and node 2 is the target, we start the traffic at node 1 and let it flow until it reaches node 2. During that time, we record a betweenness point for each node passed along the way. If the flow dead-ends before getting to node 2, we simply clear the books for that flow and restart at node 1. This is then repeated for all possible pairs, cumulating betweenness points (for successful chains only) across all pairs. Again, the justification here is that when introducing new ways of looking at centrality, it is important to begin by matching the new ways as closely as possible with existing ways.

The results show that realized betweenness in a mooch process is roughly similar to Freeman's expected values, but significant differences can be found. For example, the Strozzi family, which is ranked 7th by Freeman's measure, is found to receive far more traffic than expected. Indeed, it is ranked 2nd. The reason is that the Strozzis lie along many paths between others, even if some are not particularly short.

In the *used goods* process (trail + transfer), the traffic at any particular node randomly selects an adjacent neighbor from a list of available neighbors to move to. Neighbors are

Table 3
Freeman betweenness and frequency of arrival scores

Actor	Freeman	Package	Mooch	Used	Gossip	Infect	Money
Medici	47.5	47.5	113.7	129.8	334.3	887.03	1155.1
Guadagni	23.2	22.8	74.9	73.8	252.2	513.35	827.9
Albizzi	19.3	19.2	41.5	48.5	185.0	285.37	665.9
Salviati	13.0	13.0	26.0	26.0	168.0	182.00	503.3
Ridolfi	10.3	10.7	61.3	64.2	189.0	227.89	665.4
Bischeri	9.5	9.5	60.9	58.6	189.0	257.23	664.7
Strozzi	9.3	9.7	78.1	84.8	295.6	435.10	827.5
Barbadori	8.5	8.5	45.8	46.5	176.0	107.65	503.5
Tornabuon	8.3	8.2	58.2	59.8	189.0	222.97	666.1
Castellan	5.0	5.0	64.5	64.7	188.7	277.20	665.3
Peruzzi	2.0	2.0	59.1	55.1	189.0	232.30	664.7
Acciaiuol	0.0	0.0	0.0	0.0	0.0	0.00	176.9
Ginori	0.0	0.0	0.0	0.0	0.0	0.00	176.8
Lambertes	0.0	0.0	0.0	0.0	0.0	0.00	176.6
Pazzi	0.0	0.0	0.0	0.0	0.0	0.00	177.2

available if the undirected edge linking them to the current node has not been used before. The results closely match those of the mooch process. Interestingly, the numbers for Salviati are identical across all processes considered so far. As it happens, Salviati is located on a tendril within which the sets of geodesic paths, paths, and trails are identical, so we expect all transfer processes other than walk-based ones to have identical values for those nodes.

It is interesting to note that the relative importance of the Strozzi family is greater than expected by the Freeman measure in all flow processes other than the one Freeman's measure is designed for. In particular, Strozzi importance is particularly high in all of the trail-based flows (gift and gossip). The basic pattern seems to be that the more allowable trajectories there are, the greater the share of traffic passing through the Strozzi family. It is only when we restrict flows to traveling along shortest paths that Strozzi drops in importance.

While we do not normally interpret the absolute magnitude of centrality values, we can do it here because all of the measures measure the same thing on the same network. Thus, it is meaningful that the two serial replication processes, gossip and infection, have larger numbers in Table 3 than the transfer processes (package, mooch, and gift). This is because serial replication leads to considerably more traffic flowing through the network, since multiple copies of the same thing exist simultaneously.

4.2. Closeness centrality

A simulation of the *package delivery* process for closeness centrality is of little interest except to check the software implementation. By design, the traffic moves only along shortest paths, and therefore arrives in time proportion to the geodesic distance without variation. However, for completeness, the simulation was run and the result posted in Table 4. Since traversing one link takes one unit of time in the simulation space, the numbers in the table are identical to the Freeman closeness scores.

Table 4
Freeman closeness and first arrival times

Actor	Freeman	Package	Mooch	Used	Gossip	Infect	Money
Medici	25	25.0	46.7	50.1	78.9	63.7	575.2
Ridolfi	28	28.0	57.5	60.6	95.7	70.8	587.7
Albizzi	29	29.0	55.7	53.3	100.7	68.6	562.3
Tornabuon	29	29.0	56.4	58.1	98.2	70.0	584.8
Guadagni	30	30.0	53.7	54.8	109.3	68.8	575.3
Barbadori	32	32.0	60.5	55.3	112.3	73.1	584.4
Strozzi	32	32.0	59.9	61.3	104.0	73.3	602.9
Bischeri	35	35.0	61.1	63.9	111.6	74.1	599.0
Castellan	36	36.0	58.3	64.6	125.8	73.3	599.2
Salviati	36	36.0	57.6	59.9	94.3	72.7	533.0
Acciaiuol	38	38.0	59.5	64.3	98.2	69.8	536.3
Peruzzi	38	38.0	61.3	67.9	111.3	75.4	603.7
Ginori	42	42.0	68.9	65.3	124.5	75.9	523.2
Lambertes	43	43.0	66.4	69.8	109.6	76.1	538.2
Pazzi	49	49.0	70.7	72.9	155.9	78.8	497.8

The *mooch* simulation is of greater interest. As shown in Table 4, the average times were considerably higher for this process than for package delivery. This is because the traffic is essentially wandering randomly over the graph until it reaches the target. However, the rank-orders are quite similar.

The *gift process* shows very similar results to the mooch process, although there are some differences in the rank orders. For example, Barbadori is 4th in gift process, compared to 10th in the mooch process. The average times are only slightly higher than for the mooch process, which is surprising, since trails can be much longer than paths (trails are limited by the number of edges in a graph whereas paths are limited by the number of nodes).

It should be noted that Table 4 gives times to arrive at the focal node rather than times it takes to reach other nodes from the focal node. In an undirected graph such as we have here, we normally expect closeness to be symmetrical, since the distances from a node are the same as the distances to a node. However, for traffic that does not flow via shortest paths, distances can be non-symmetric even in undirected graphs. For example, consider a pair of adjacent nodes, one with degree 10 and one with degree 1 (a pendant). Assume a process in which at time *t*, a node selects one alter at random and sends traffic to that node. When the process begins with the pendant, it unfailingly arrives at the degree 10 node in one unit of time. However, when the process begins with degree 10 node, it reaches the pendant in one unit of time only 1 in 10 times—all other times it arrives later. For a transfer process, on average, it will arrive in five units of time. Table 5 gives times from the focal node to all others.

Comparing with Table 4, one can see that the in and out figures are extremely close for all nodes except Ginori, which in most processes can reach others more easily than others can reach it. This is not a statistical fluke. For random walks, the average distance between two nodes depends on the ordering of opportunities for paths that are encountered. For example, consider the link between Ginori and Albizzi. In an infection process, an infection originating with Ginori reaches Albizzi in one unit of time every time. But an infection

Table 5
Freeman closeness and first reach times

Actor	Freeman	Package	Mooch	Used	Gossip	Infect	Money
Medici	25	25.0	43.9	42.7	102.4	40.4	172.5
Ridolfi	28	28.0	54.9	54.5	108.8	58.6	287.0
Albizzi	29	29.0	57.3	58.2	107.6	61.8	387.9
Tornabuon	29	29.0	55.0	54.5	107.1	58.7	292.3
Guadagni	30	30.0	54.8	53.5	105.0	52.8	285.0
Barbadori	32	32.0	63.0	60.1	112.6	69.4	446.0
Strozzi	32	32.0	52.0	53.0	108.6	58.8	296.1
Bischeri	35	35.0	57.9	59.0	110.1	62.5	350.6
Castellan	36	36.0	57.3	57.9	110.8	73.4	390.8
Salviati	36	36.0	57.7	67.2	103.7	85.8	777.4
Acciaiuol	38	38.0	59.8	71.4	107.6	90.4	870.0
Peruzzi	38	38.0	59.4	60.2	112.1	74.2	408.5
Ginori	42	42.0	78.1	72.0	114.0	104.0	1084.5
Lambertes	43	43.0	72.0	77.7	110.7	94.3	981.2
Pazzi	49	49.0	70.9	80.1	109.3	99.2	1473.2

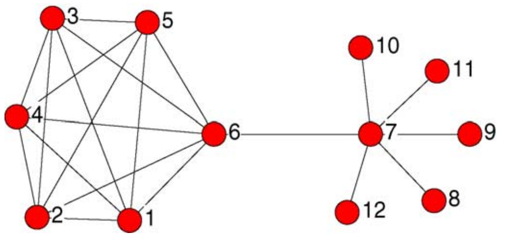

Fig. 5. In the random walk of a mooch process, nodes on the right will reach the nodes on the left more quickly than the other way around.

originating with Albizzi reaches Ginori on average in two units of time. This is because Albizzi has three possibilities, so Ginori could either be first, in which case the infection arrives in one unit of time, or second, in which case it arrives in two units, or third, in which case it arrives in three units. The average is two.

Similarly, in Fig. 5, nodes 1–5 take an average of 5.8 links to reach nodes 8–12. In contrast, nodes 8–12 take only 5.0 links to reach nodes 1–5. The reason is that when members of a populous clique begin a random walk, they are very likely to spend quite a bit of time wandering around their own clique. Few available paths at any juncture actually lead out of the clique. In contrast, the peripheral members of a star are funneled out of the star and quickly enter the clique, in which all paths lead to all nodes in relatively short order.

5. Discussion

A common criticism of social network research is that insufficient attention is paid to network dynamics (e.g., Watts, 2003). In the case of assessing nodal prominence (i.e., centrality), the critics are right. As Friedkin (1991) has noted, the discussion of centrality has largely avoided any mention of the dynamic processes that unfold along the links of a network (not to mention the processes that shape the network structure). Yet, the importance of a node in a network cannot be determined without reference to how traffic flows through the network. The simulation studies reported in this paper show that nodes that are highly central in trail-based processes such as gossip flow need not be highly central in, say, geodesic-based processes such as package delivery. The characteristics of the flow process affect which nodes will receive flows (quickly, frequently, and certainly) and which are in a position to control flows. Just as researchers in experimental exchange theory have investigated how network position interacts with the rules of the game to generate opportunities for power use, I have sought to explore how node importance results from an interaction between position and the characteristics of the flow process. To this end, this paper offers an analysis of flow processes, noting key dimensions along which they differ from each other and constructing a simple typology of processes.

In the past, centrality has been considered an abstract property of a node's position in a network, measurable without regard for what the nodes and links mean and what processes they might support. No doubt many authors of centrality measures did not intend their measures to serve as theoretical predictive models. Yet, the measures certainly embody theoretical thinking about network phenomena. In this paper, I propose an alternative conception that views centrality as a node-level outcome of implicit models of flow processes. More specifically, I regard the formulas for centrality concepts like betweenness and closeness as generating the expected values – under specific unstated flow models – of certain kinds of node participation in network flows. As such, they do not actually measure node participation at all but rather indicate the expected participation if things flow in the assumed way. One contribution of this paper is to make explicit what the assumptions behind each measure are, and then to test this deconstruction via simulation. For example, when we measure how often a node handles a package in a package delivery process, the results match to a few decimal places what the Freeman betweenness formula predicts.

Thinking of existing centrality measures as models begs the important question of what exactly they are models of. To say that centrality measures measure "node importance" or "node participation" is not specific enough, since there are multiple measures that make the same flow assumptions (e.g., Freeman's closeness and betweenness measures). For closeness and betweenness centrality, there are clear answers. In the context of network flow, the essence of closeness is time-until-arrival of something flowing through the network. The Freeman formula provides expected values of arrival times for package deliveries and other flow processes in traffic moves along shortest paths or take all paths simultaneously. In contrast, the essence of betweenness is frequency of arrival. The Freeman formula provides expected values for how often packages pass a station in a package delivery system.[5] Thus, an objective of this paper has been to put forth the notion that the essence of measures like closeness and betweenness can be separated from the particulars of their formulas which embody the characteristics of the flow processes for which they were designed. A complete typology of centrality measures would therefore include not only the dimensions pertaining to flow characteristics, as in Table 2, but also to the aspect of node participation captured (such as first arrival time and arrival frequency).

As noted earlier, a striking thing about the set of centrality measures currently in existence is the absence of measures designed for the flow processes of greatest interest. The Freeman measures which dominate empirical network analysis are largely misapplied, since the processes of interest are typically not based on geodesic paths. Thus, there is a real need for new measures that apply to more realistic flow processes. Of course, as this paper has shown, we can use simulation to obtain estimates of the expected values for any flow process. However, simulations are relatively costly and are not suitable for large graphs. Therefore, a crucial next step is the development of analytical solutions – i.e., formulas – for the expected values for arrival times and frequencies for a variety of different flow models, a task that Friedkin (1991) and Newman (in press) have already begun.

Finally, it should be noted as a limitation of this study that I have only considered the case of flows that have a source and a target. This was done in order to ensure the comparability of existing centrality formulas with the results generated from flow simulations and establish the validity of my approach. However, the underlying logic of some of these flows (which are essentially random traversals of a network) suggests that we should also examine the case

[5] It is worth noting in passing that time until first arrival and frequency of arrival are concepts that are well studied in the context of Markov processes and provide a bridge to that literature.

where flows originate at each node systematically, but have no particular target. This will pose some challenges for walk-based processes but is an important line of future research.

Acknowledgments

I am grateful to Mark Newman for help and encouragement, along with Phil Bonacich and Martin Everett. This research was supported by Office of Naval Research grant number N000140211032.

References

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2, 113–120.
Bonacich, P., 1987. Power and centrality: a family of measures. American Journal of Sociology 92, 1170–1182.
Bonacich, P., 1991. Simultaneous group and individual centralities. Social Networks 13, 155–168.
Borgatti, S.P., 1995. Centrality and AIDS. Connections 18 (1), 112–114.
Coleman, J.S., Katz, E., Menzel, H., 1966. Medical Innovation: A Diffusion Study. Bobbs-Merrill, Indianapolis.
Freeman, L.C., 1979. Centrality in networks: I. Conceptual clarification. Social Networks 1, 215–239.
Freeman, L.C., Borgatti, S.P., White, D.R., 1991. Centrality in valued graphs: a measure of betweenness based on network flow. Social Networks 13, 141–154.
Friedkin, N.E., 1991. Theoretical foundations for centrality measures. American Journal of Sociology 96, 1478–1504.
Hoede, C. 1978. A new status score for actors in a social network. Department of Mathematics, Twente University, unpublished manuscript.
Hubbell, C.H., 1965. An input–output approach to clique identification. Sociometry 28, 377–399.
Katz, L., 1953. A new index derived from sociometric data analysis. Psychometrika 18, 39–43.
Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.
Newman, M.E.J. A measure of betweenness centrality based on random walks. Social Networks, in press.
Noh, J.D., Rieger, H., 2004. Random walks on complex networks. Physical Review Letters 92, 118701-1-4.
Padgett, J.F., Ansell, C.K., 1993. Robust action and the rise of the Medici, 1400–1434. American Journal of Sociology 98, 1259–1319.
Stephenson, K., Zelen, M., 1989. Rethinking centrality: methods and examples. Social Networks 11, 1–37.
Taylor, M., 1969. Influence structures. Sociometry 32, 490–502.
Watts, D., 2003. Six Degrees: The Science of a Connected Age. Norton, New York.