C:\Papers\Probabilistic Information Flow\Centrality and Network Flow 3b.doc

# Centrality and Network Flow[♣]

Stephen P. Borgatti
Dept. of Organization Studies
Boston College

# EARLY DRAFT! Contains Errors!

**Abstract**

Centrality measures, or at least our interpretations of these measures, make implicit assumptions about the manner in which things flow 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 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 things flow.

## 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,

[♣] 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. 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.

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 eigenvector centrality or Katz's influence measure (Katz, 1953), count walks, which assume that flows can not only be circuitous, but revisit nodes and lines at will 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 (like when we compute the mean of a nominal-scaled variable) or we get poor answers (like when we use regression to predict values of a dependent variable, but assume a linear model 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.

More generally, the aim of this paper is to recast centrality measures as generating the expected values based on implicit theoretical models of network flows. This enables us to test the models empirically, as well as to apply them in a more thoughtful way.

## 2. Typology of Flow Processes

In this section I consider a number of commonly encountered flow processes. It is not necessary for my purposes 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 other flows. I consider each of the following different kinds of flows: gifts, currency, transportation, delivery, gossip, e-mail, infection, and influence.

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

*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 place at one 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, then if B receives the book again, he or she does not send it 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.

*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.

*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 don't 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.

*E-mail.* A typical example is an email message that warns of an electronic virus. The message is forwarded from one person to several of his contacts, often by sending to all of them simultaneously (unlike confidential gossip). The message exists in multiple places at the same time, thanks to diffusion by replication.

*Attitudes*. Here the notion is of an influence process in which, through interaction, individuals effect changes in each others beliefs or attitudes. Thus, attitudes about what fashion items are cool versus passé are spread from person to person. The attitudes spread through replication rather than transfer (I don't 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.

*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 doesn't re-infect anyone who already has had it because they become immune.

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

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) or simultaneous, like a broadcast (parallel). A third attribute concerns whether the traffic flows non-deterministically, meaning that any particular juncture it is always able to take the "best" way (such as taking the shortest possible road to a predetermined target), or is simply flowing 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 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 could be 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..

**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 | Markov |

Of course, there are other dimensions we can discern that distinguish among flow processes. For example, some processes might always involve a 2-way effect in which A influences B at the same time that B influences A. However, since this is new ground, it is better to start by keeping it simple and seeing where it takes us.

## 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 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-position to obtain novel information early, when it has the most value. Thus, organizations with low closeness in 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 see who is well positioned 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 assumption of shortest paths 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 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 is basically a measure of the number 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 possible path. 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 Figure 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 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.

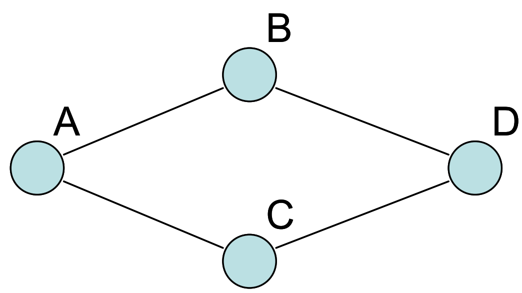


Figure 1. Traffic flowing from A to D is expected to through B or through C with equal probability.

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. Rather, the assumptions built in to the 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

$$\lambda v = Av$$

where A is the adjacency matrix of the graph, lambda is a constant (the eigenvalue) and 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, Hubbell, Taylor, Hoede, Coleman and Friedkin, almost all of which are known as influence measures. The idea is that even if a node influences just one other node, if that node influences many others (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; 1992) that an eigenvector is proportional to the row sums of a matrix 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.

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

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 to see 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 see 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 something in the next time period 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.[3] 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.

Table 2 locates the four best known measures of centrality in the boxes of the typology for which they are appropriate. A striking fact is that not only is much of the box empty, but most of the sociologically interesting processes are not covered by the major measures. However, if we consider less well-known and/or forthcoming measures, the situation improves somewhat. For example, the influence and status measures of Katz (), Hubbell (), Taylor () and Hoede () all fall in the [walks, parallel] cell of the table, along with the closeness measure of Stephenson and Zelen (1989). In contrast, the random walk betweenness and closeness measures of Newman (2004) and Noh and Rieger (2003), respectively, fall into the [walks,transfer] cell of the table. Similarly, Friedkin's (1991) measures of influence also fall in that cell.

**Table 2**
**Flow Processes and Major Centrality Measures**

| | parallel duplication | serial duplication | transfer |
|---|---|---|---|
| geodesics | | - Freeman closeness | - Freeman closeness<br>- Freeman betweenness |
| paths | - Freeman closeness<br>- Freeman degree | | |

[3] 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 $\pi$ are based on degree, than $\pi_i = d_i/\sum d_i$. We then need to show that $\pi P = \pi$. Writing $\pi P$ as $\sum_i \pi_i p_{ij}$ we can then substitute $d_i/\sum d_i$ for $\pi_i$ 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 $\pi_j$, the result we were looking for.

| | | | |
|---|---|---|---|
| trails | - Freeman closeness<br>- Freeman degree | | |
| walks | - Freeman closeness<br>- Freeman degree<br>- Bonacich eigenvector | | |

## 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 badly 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 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 following algorithmic structure:

```
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
```

Figure 1. Source/Target method of running simulations

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 following structure:

```
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
```

Figure 2. Source-only method of running simulations.

However, for the purpose of maintaining comparability when validating and analyzing Freeman's betweenness measure, 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 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 the discussion section.

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 is used to illustrate the simulations. The network is shown in Figure xx. 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.

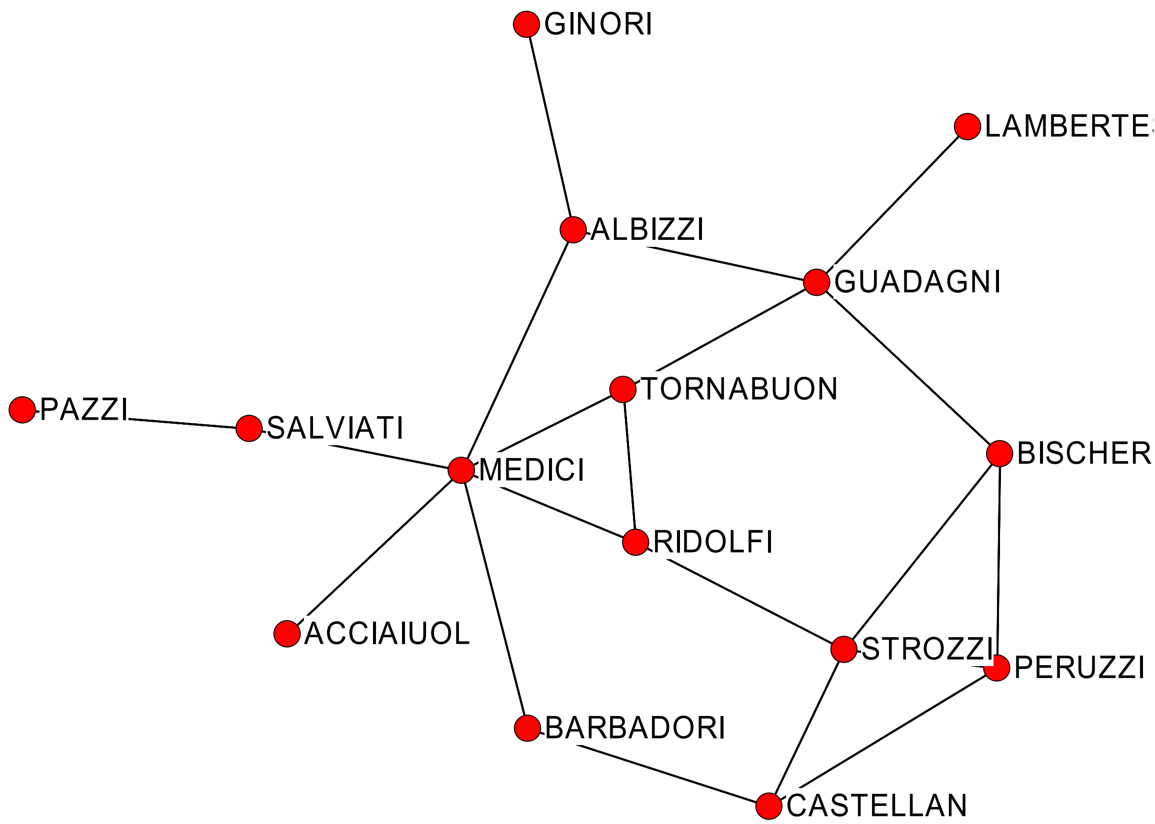


Figure 3. Padgett's data on marriage ties among Renaissance Florentine families (isolate removed).

*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 – henceforth "(geodesic+transfer)". 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 of F, S and T are systematically tried. Finally, each run is repeated 1,000 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 compares the results of 1,000 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.

**Table 3**
**Freeman and realized betweenness scores**

| Node | Freeman | Package | Mooch | Used Goods | Gossip | Infection |
|---|---|---|---|---|---|---|
| MEDICI | 47.5 | 47.5 | 56.8 | 64.9 | 167.1 | 443.5 |
| GUADAGNI | 23.2 | 22.6 | 37.5 | 36.9 | 126.1 | 256.7 |
| ALBIZZI | 19.3 | 19.1 | 20.8 | 24.2 | 92.5 | 142.7 |
| SALVIATI | 13.0 | 13.0 | 13.0 | 13.0 | 84.0 | 91.0 |
| RIDOLFI | 10.3 | 10.8 | 30.6 | 32.1 | 94.5 | 113.9 |
| BISCHERI | 9.5 | 9.5 | 30.4 | 29.3 | 94.5 | 128.6 |
| STROZZI | 9.3 | 9.9 | 39.0 | 42.4 | 147.8 | 217.5 |
| BARBADORI | 8.5 | 8.5 | 22.9 | 23.2 | 88.0 | 53.8 |
| TORNABUON | 8.3 | 8.0 | 29.1 | 29.9 | 94.5 | 111.5 |
| CASTELLAN | 5.0 | 5.0 | 32.2 | 32.4 | 94.3 | 138.6 |
| PERUZZI | 2.0 | 2.0 | 29.5 | 27.5 | 94.5 | 116.1 |
| ACCIAIUOL | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GINORI | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| LAMBERTES | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| PAZZI | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

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 comparing as closely as possible with existing ways.

The results show that realized betweenness in a mooch process are roughly similar to Freeman's expected values, but significant differences can be found. For example, the Strozzi family, which is ranked 7$^{th}$ by Freeman's measure, is found to receive far more traffic than expected. Indeed, it is ranked 2$^{nd}$. The reason is that the Strozzi's 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 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.

The

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 don't 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.

### *4.1. 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 anyway 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 and realized in-closeness scores**

| Actor | Freeman | Package | MoochInClo | GiftInClo | GossipIn | InfInDist |
|---|---|---|---|---|---|---|
| MEDICI | 25 | 25.0 | 46.7 | 50.1 | 78.9 | 40.4 |
| RIDOLFI | 28 | 28.0 | 57.5 | 60.6 | 95.7 | 58.6 |
| ALBIZZI | 29 | 29.0 | 55.7 | 53.3 | 100.7 | 61.8 |
| TORNABUON | 29 | 29.0 | 56.4 | 58.1 | 98.2 | 58.7 |
| GUADAGNI | 30 | 30.0 | 53.7 | 54.8 | 109.3 | 52.8 |
| BARBADORI | 32 | 32.0 | 60.5 | 55.3 | 112.3 | 69.4 |
| STROZZI | 32 | 32.0 | 59.9 | 61.3 | 104.0 | 58.8 |
| BISCHERI | 35 | 35.0 | 61.1 | 63.9 | 111.6 | 62.5 |
| CASTELLAN | 36 | 36.0 | 58.3 | 64.6 | 125.8 | 73.4 |
| SALVIATI | 36 | 36.0 | 57.6 | 59.9 | 94.3 | 85.8 |
| ACCIAIUOL | 38 | 38.0 | 59.5 | 64.3 | 98.2 | 90.4 |
| PERUZZI | 38 | 38.0 | 61.3 | 67.9 | 111.3 | 74.2 |
| GINORI | 42 | 42.0 | 68.9 | 65.3 | 124.5 | 104.0 |
| LAMBERTES | 43 | 43.0 | 66.4 | 69.8 | 109.6 | 94.3 |
| PAZZI | 49 | 49.0 | 70.7 | 72.9 | 155.9 | 99.2 |

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 fourth in gift process, compared to 10$^{th}$ 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 reach the focal node rather than times from the focal node to all others. 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 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 one in ten times – all other times it arrives later. For a transfer process, on average, it will arrive in 5 units of time. Table 5 gives times from the focal node to all others.

**Table 5**
**Freeman and realized out-closeness scores**

| Actor | Freeman | Package | MoochOutClo | GiftOutClo | GossipOut | InfOutDist |
|---|---|---|---|---|---|---|
| MEDICI | 25 | 25.0 | 43.9 | 42.7 | 102.4 | 63.7 |
| RIDOLFI | 28 | 28.0 | 54.9 | 54.5 | 108.8 | 70.8 |
| ALBIZZI | 29 | 29.0 | 57.3 | 58.2 | 107.6 | 68.6 |
| TORNABUON | 29 | 29.0 | 55.0 | 54.5 | 107.1 | 70.0 |
| GUADAGNI | 30 | 30.0 | 54.8 | 53.5 | 105.0 | 68.8 |
| BARBADORI | 32 | 32.0 | 63.0 | 60.1 | 112.6 | 73.1 |
| STROZZI | 32 | 32.0 | 52.0 | 53.0 | 108.6 | 73.3 |
| BISCHERI | 35 | 35.0 | 57.9 | 59.0 | 110.1 | 74.1 |
| CASTELLAN | 36 | 36.0 | 57.3 | 57.9 | 110.8 | 73.3 |
| SALVIATI | 36 | 36.0 | 57.7 | 67.2 | 103.7 | 72.7 |
| ACCIAIUOL | 38 | 38.0 | 59.8 | 71.4 | 107.6 | 69.8 |
| PERUZZI | 38 | 38.0 | 59.4 | 60.2 | 112.1 | 75.4 |
| GINORI | 42 | 42.0 | 78.1 | 72.0 | 114.0 | 75.9 |
| LAMBERTES | 43 | 43.0 | 72.0 | 77.7 | 110.7 | 76.1 |
| PAZZI | 49 | 49.0 | 70.9 | 80.1 | 109.3 | 78.8 |

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 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, for an infection process running in the graph in Figure 4, nodes 1 through 5 take an average of 8 units of time to reach nodes 8 through 12, as shown in Table 6. In contrast, nodes 8 through 12 take only 7.0 units to reach nodes 1 through 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 start are funneled out of the star and quickly enter the clique, in which all paths lead to all nodes in relatively short order.

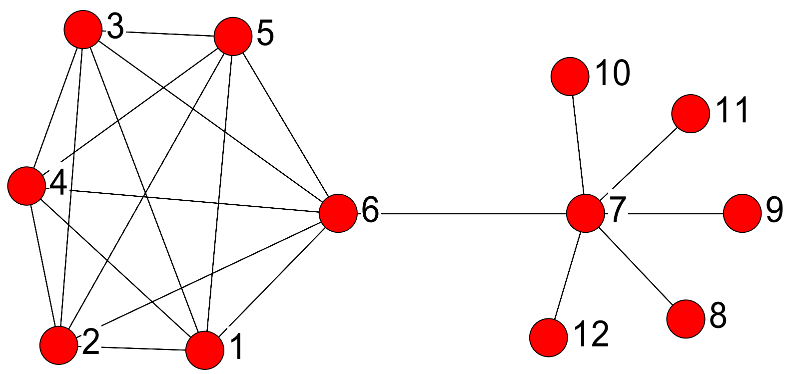


Figure 4. In the random walk of a Mooch process, nodes on the right will reach the nodes on the left more quickly than the other way around.

Similarly, in Figure 4, nodes 1 through 5 take an average of 5.8 links to reach nodes 8 through 12, as shown in Table 6. In contrast, nodes 8 through 12 take only 5.0 links to reach nodes 1 through 5. Similarly, nodes 1 through 5 require an average 2.8 links to reach themselves, while nodes 8 through 12 require only 2.0 links. 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 start are funneled out of the star and quickly enter the clique, in which all paths lead to all nodes in relatively short order.

Table 6
Travel times between nodes in Figure 4

| | 1-5 | 8-12 |
|---|---|---|
| 1-5 | 2.8 | 5.8 |
| 8-12 | 5.0 | 2.0 |

## 5. Discussion

The main point of this article is to call attention to the assumptions made by off-the-shelf centrality measures about the ways things flow through a network. This is one case where critics of social network research claiming that insufficient attention is paid to network dynamics are right. Structural position is not the only determinant node importance. Node importance results from an interaction between position and the characteristics of the flow process, such as whether traffic moves from node to node or is copied, and whether the trajectories followed resemble graph theoretic paths, trails or walks. The simulation studies reported here bear this out. For example, when we measure observed betweenness in a trail-based process, the results are quite different from the expected values generated by Freeman's betweenness centrality formula. When we measure observed betweenness in a package delivery process, the results are virtually identical to what the formula predicts.

Of course, it could be argued that centrality measures were not intended to serve as predictive models. From a historical point of view, this may be true. However, part of my purpose here is precisely to suggest that centrality measures can and should be thought of as models. Certainly they embody – some more explicitly than others – theoretical thinking about network phenomena. Taking this thinking to its conclusion and conceiving of centrality measures as generating expected values from a model opens the possibility of statistical testing of network dynamics. For example, if we obtain empirical data on network links and on time until arrival of information, we can statistically test a hypothesis about the characteristics of the latent flow process that was operating.

Thinking of centrality measures as models begs the important question of what exactly they are models of. To say "node importance" is not specific enough, and makes centrality something you define rather than measure. For closeness and betweenness centrality, I believe 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 which things move 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. For many other measures, what exactly they model may not be easy to divine. From the perspective taken here, this is an indictment of such measures – it is another way of saying that we don't know what substantively they measure.

As noted earlier, a striking thing about Table 2 is the absence of measures available for most flow processes of real interest. In the absence of computational formulas for these, we can of course use the simulations to generate nearly exact values. 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 under different flow models with respect to arrival frequencies and times, corresponding to the betweenness and closeness aspects of centrality, respectively.

Implicit in this paper has been 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. Thus, notions like closeness and betweenness are general aspects of centrality that are meaningful in any kind of flow process (if properly measured) and are conceptually distinct from each other. Thus, there are two fundamental ways in which centrality measures differ from each other (in addition to myriad smaller differences). These are flow assumptions and aspect of centrality. Thus a pair of centrality measures like Newman's random walk betweenness and Freeman et al's flow betweenness differ on flow assumptions but capture the same aspect of centrality. In contrast Freeman's closeness and Freeman's betweenness can be applied to the same flows, but measure different aspects of centrality. This paper has only dealt with two aspects of centrality – closeness and betweenness. However, it should not be assumed that these are the only aspects of centrality. What is needed is an analysis of the full range of proposed centrality measures to understand which represent measurements of fundamentally different aspects of centrality and which do not. This too, then, provides direction for future next steps.

Finally, in order to ensure the comparability of centrality formulas with the results generated from flow simulations, I have only considered the case of flows that have a source and a target. 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 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 research.

## 6. References


Alexander, C.N. 1963. "A Method for Processing Sociometric Data." *Sociometry* 26:268-9.

Bavelas, A. 1948. "A Mathematical Model for Group Structure." *Human Organization* 7:16-30.

Bavelas, A. 1950. "Communication Patterns in Task Orientated Groups." *Journal of the Acoustical Society of America* 22:271-88.

Blau, P.M. 1963. "*The Dynamics of Bureaucracy: A Study of Interpersonal Relations in Two Government Agencies*: Chicago: Univ. of Chicago Press.

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* 13:155-68.

Borgatti, S. P. 1995. Centrality and AIDS. *Connections*, 18(1): 112-114Burt, R.S. 1982. *Toward a Structural Theory of Action*. New York: Academic Press.

Doreian, P. 1986. "Measuring Relative Standing in Small Groups and Bounded Social Networks." *Social Psychological Quarterly* 49:247-59.

Ford, L.R. and D.R. Fulkerson. 1956. "Maximal Flow Through a Network." *Canadian Journal of Mathematics* 8:399-404.

Ford, L.R. and D.R. Fulkerson. 1962. *Flows in Networks*. Princeton: Princeton University Press.

Freeman, L.C. 1979. "Centrality in Networks: I. Conceptual Clarification." *Social Networks* 1:215-39.

Freeman, L.C. 1980. "The Gatekeeper, Pair-Dependency, and Structural Centrality." *Quality and Quantity* 14:585-92.

Freeman, L.C. 1983. "Spheres, Cubes and Boxes: Graph Dimensionality and Network Structure." *Social Networks* 5:139-56.

Freeman, L.C., S.P. Borgatti, and D.R. White. 1991. "Centrality in Valued Graphs: A Measure of Betweenness Based on Network Flow." *Social Networks* 13:141-54.

Friedkin, N.E. 1991. "Theoretical Foundations for Centrality Measures." *American Journal of Sociology* 96:1478-504.

Galaskiewicz, J. 1979. *Exchange Networks and Community Politics*. Beverly Hills: Sage Publications.

Gould, R.V., and R.M. Fernandez. 1989. "Structures of Mediation: A Formal Approach to Brokerage in Transaction Networks." Pp. 89-126 in *Sociological Methodology 1989* edited by C.C. Clogg. Ann Arbor, MI: Blackwell.

Granovetter, M.S. 1974. *Getting a Job: A Study in Contacts and Careers*. Cambridge, MA: Harvard University Press.

Harary, F. 1969. *Graph Theory*. Reading, Mass: Addison-Wesley.

Hoede, C. 1978. "A New Status Score for Actors in a Social Network." unpublished manuscript, dept. of mathematics, Twente University.

Hubbell, C.H. 1965. "An Input-Output Approach to Clique Identification." *Sociometry* 28:377-99.

Johnson, S.C. 1967. "Hierarchical Clustering Schemes." *Psychometrika* 32:241-54.

Katz, L. 1953. "A New Index Derived From Sociometric Data Analysis." *Psychometrika* 18:39-43.

Knoke, D. and R.S. Burt. 1983. "Prominence." Pp. 195-222 in *Applied Network Analysis: A Methodological Introduction* edited by R.S. Burt and M. Minor. Beverly Hills: Sage Publications.

Laumann, E.O. and F.U. Pappi. 1973. "New Directions in the Study of Community Elites." *American Sociological Review* 38:212-30.

Laumann, E.O. and F.U. Pappi. 1976. *Networks of Collective Action*. New York: Academic Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Leavitt, H.J. 1951. "Some Effects of Certain Communication Patterns on Group Performance." *Journal of Abnormal and Social Psychology* 46:38-50.

Mariolis, P. 1975. "Interlocking Directorates and Control of Corporations." *Social Science Quarterly* 56:425-39

Markovsky, B., D. Willer and T. Patton. 1988. "Power Relations in Exchange Networks." *American Sociological Review* 53:220-236.

Marsden, P.V. 1982. "Brokerage Behavior in Restricted Exchange Networks." Pp. 201-18 in *Social Structure and Network Analysis*, edited by P.V. Marsden and N. Lin. Beverly Hills: Sage Publications.

Marsden, P.V. and E.O. Laumann. 1977. "Collective Action in a Community Elite: Exchange, Influence Resources, and Issue Resolution." Pp. 199-250 in *Power, Paradigms and Community Research*, edited by R.J. Liebert and A.W. Imershein. Beverly Hills: Sage Publications.

Menger, K. 1927. "Zur Allgemeinen Kurventheorie." *Fund. Math.* 10:96-115.

Mintz, B. and M. Schwartz. 1981a. "Interlocking Directorates and Interest Group Formation." *American Sociological Review*. 46:851-68.

Mintz, B. and M. Schwartz. 1981b. "The Structure of Intercorporate Unity in American Business." *Social Problems*. 28:87-103.

Mintz, B. and M. Schwartz. 1985. *The Power Structure of American Business*. Chicago: University of Chicago Press.

Mizruchi, M. 1982. *The Structure of the American Corporate Network: 1904-1974*. Beverly Hills: Sage.

Mizruchi, M.S., P. Mariolis, M. Schwartz, B. Mintz. 1986. "Techniques for Disaggregating Centrality Scores in Social Networks." Pp. 26-48 in *Sociological Methodology 1986* edited by N.B. Tuma. San Francisco: Jossey-Bass.

Nakao, K. 1990. "Distribution of Measures of Centrality: Enumerated Distributions of Freeman's Graph Centrality Measures." *Connections* 13:10-22.

Nemeth, R. and D.A. Smith. 1985. "International Trade and World-System Structure: A Multiple Network Analysis." *Review* 8:517-60.

Nunnally, J.C. 1967. *Psychometric Theory*. New York: McGraw Hill.

Padgett, J. F. and Ansell, C. K., 1993. Robust action and the rise of the Medici, 1400-1434. Am. J. Sociol. 98, 1259-1319.

Pitts, F.R. 1979. "The Medieval River Trade Network of Russia Revisited." *Social Networks* 1:285-92.

Sabidussi, G. 1966. "The Centrality Index of a Graph." *Psychometrika* 31:581-603.

Sade, D.S. 1972. "Sociometrics of *Macaca Mulatta* I: Linkages and Cliques in Grooming Networks." *Folia Primatologica* 18:196-223.

Sade, D.S. 1989. "Sociometrics of *Macaca Mulatta* III: N-Path Centrality in Grooming Networks." *Social Networks* 11:273-92.

Stephenson, K. and M. Zelen. 1989. "Rethinking Centrality: Methods and Examples." *Social Networks* 11:1-37.

Taylor, M. 1969. "Influence Structures." *Sociometry* 32:490-502.