2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust

Attentive Betweenness Centrality (ABC): Considering Options and Bandwidth when Measuring Criticality

Sibel Adalı
Computer Science Department
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: sibel@cs.rpi.edu

Xiaohui Lu
Computer Science Department
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: lux3@cs.rpi.edu

Malik Magdon-Ismail
Computer Science Department
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: magdon@cs.rpi.edu

***Abstract*—Betweenness centrality measures how critical a node is to information flow in a network. A node is critical (and hence should have high betweeness) if it is on many shortest paths. Two shortcomings of such a measure are:**

(i) It ignores nodes on "almost shortest" paths;
(ii) It assumes that a node can provide the same attention to information flow through each of those shortest paths, no matter how many shortest paths the node controls.

There have been attempts to address these concerns in the literature, with partial success. We provide a new measure, *attentive betweenness centrality (ABC)*, that measures criticality by the amount of attention a node devotes to the information flow between other nodes. Our measure addresses both the aforementioned concerns and can be computed efficiently. It performs as well or better than betweenness centrality on both stylized networks and large scale real data networks, and hence provides a useful tool for measuring node criticality.

I. Introduction

Betweenness is a measure of how critical a node is in a network. It is one of a number of 'centrality' indices that have been introduced by researchers over the years (others include closeness centrality, degree centrality, ..., [1]). Such indices have proved invaluable in understanding the roles of actors in social networks, and more generally the importance of vertices in information networks, citation networks, computer and communication networks, biological networks, etc.

Even within the realm of betweenness, there are several variants, and the standard betweenness centrality (or more simply just betweenness) measures the fraction of shortest paths that pass through a node [2], [3]. It models how critical that node is in the transfer of information between other pairs of nodes in the network. Bridge nodes in a network tend to have high betweenness. Betweenness can be computed for all nodes in a network with m edges and n nodes in $O(mn)$ [4].

Betweenness rewards a node if it is on a shortest path, and does not reward it at all if it is not on a shortest path. This is a non-intuitive approach given that betweenness is supposed to capture how critical a node is to *information flow*, because in most networks, information does not only flow along shortest paths [5], [6]. In fact, when actors are propagating information, there is no reason to expect that they even know what the most direct path to the destination is, as was demonstrated in [7], [8]. So, on the philosophical side, if betweenness is to capture how much control an actor has over the information flow, it should *not* be assumed that information has to flow along shortest paths. In fact, assuming that information must flow only along shortest paths leads to undesirable consequences, like overweighting the criticality of nodes that happen to be on *the* shortest path while completely marginalizing nodes who happen to be on paths that are just a little longer. The following example in Figure 1 illustrates. One can easily verify that

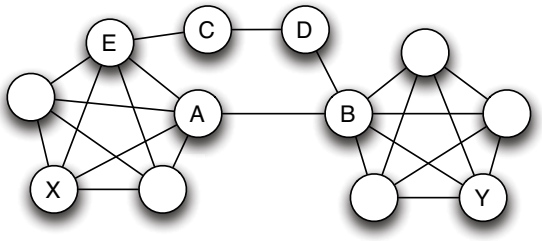


Fig. 1. Stylized network to illustrate the effect of ignoring nodes not on the shortest path.

nodes C and D are not on the shortest path between any other pair of nodes with one in the left clique and the other in the right clique, while A and B are on all the shortest paths from the left clique to the right one. Hence, the betweenness of A and B is high, while the betweenness of C and D is very low; this does not seem to be a reasonable conclusion. Even more, it is clear that B is more important than A, since without B there are no paths from the left to the right, but the reverse is not true for A. While betweenness reflects this somewhat, the difference is not appropriately emphasized. The culprit that led to this state is the ignoring of information flow along multiple paths, even if they are a little longer. Several attempts to address this concern have been proposed and we will mention the two perhaps most common approaches. The first is the flow based approach as in *flow betweenness* [6], which

978-0-7695-4848-7/12 $26.00 © 2012 IEEE
DOI 10.1109/SocialCom-PASSAT.2012.53

considers sending the maximum flow of information between two nodes and determines the flow betweenness of a third node v by what fraction of that maximum flow passes through v. The main concerns with flow betweenness are: it computes a node's importance assuming that other nodes send information along max-flow paths, and it is by no means clear that nodes could even compute max-flow paths; max-flow paths can easily over-emphasize highly indirect paths; max flow paths can completely ignore nodes which would certainly be expected to play a role in information flow (for example node C in network 2 of Figure 5 will not be used in a maximum flow from the left to the right because all such flows use the parallel paths passing through A and B); the computation of flow betweenness is $O(m^2n)$ [9] which is not scalable to modern large networks. The second general approach to including non-shortest paths in the centrality measure is based on random walks, for example *random walk betweenness* [10], which computes the betweenness of a node u by choosing the start and end point of a random walk uniformly, and computing the fraction of times this random walk will pass through u. While "all" paths are considered as possible information flow paths, such methods have a tendency to over emphasize peripheral nodes. For example, in Figure 1, the nodes in the left and right cliques get visited often because they are well connected. Hence, these 'peripheral' nodes will get high random walk betweenness, yet they clearly don't serve any critical role in the network's information flow. Further the computation time is $O(mn^2)$, which again is not scalable in comparison to betweenness.

Attention. To illustrate the concept of attention, we consider Figure 2. Let's consider information flow between A and

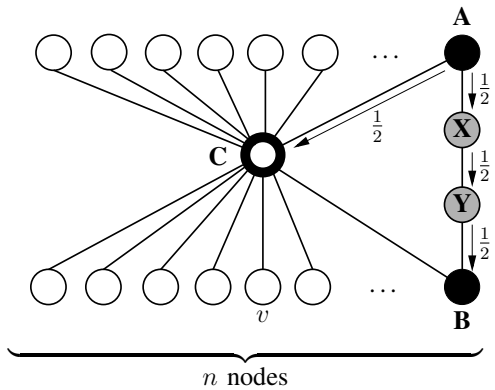


Fig. 2. Stylized network to illustrate attention.

B (shaded black) and let's analyze the criticality of nodes C, X, Y for this flow. Betweenness says X and Y are useless, because the only shortest path from A to B is through C. Flow betweenness will send equal flow down the A–X–Y–B path as the A–C–B path, and so C, X, Y all have equal flow betweenness. A random walk starting at A and ending at B can be viewed as a fair gamblers ruin problem; after a little algebra one finds that the fraction of such random walks that use C, X, Y are respectively $\frac{3}{4}, \frac{2}{3}\cdot\frac{1}{2}$. To summarize:

	actor		
Measure	C	X	Y
Betweenness	1	0	0
Flow	1	1	1
Random Walk	1	$\frac{8}{9}$	$\frac{2}{3}$

(We rescaled all the scores in each row so that the maximum is 1 for easy comparison of the relative rankings.) The general conclusion is that C is important, yet when we step back and take stock, this seems very surprising. If asked which is the more *reliable* path for the A–B information flow, most would not disagree if A would choose A–X–Y–B, and any of a number of explanations would be convincing. We choose the explanation that C is overloaded; C has only a finite attention, and an incoming piece of information will not be forwarded on to B necessarily. C has so many options, that (say) picking one at random will not get it to B with very high probability. In fact there is only a 1 in $2n-1$ chance that it gets to B. Since A, *a priori* does not know which path is better, it splits its information along both its outgoing options. Of the $\frac{1}{2}$ that goes to X, there is nowhere to go but B. Of the $\frac{1}{2}$ that goes to C, $1/(2n-1)$ of that will reach B (assuming C randomizes). The total flow coming to B is therefore

$$\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2n-1}$$

Of this information reaching B, the fraction that came from X, Y is $1 - \frac{1}{2n}$ and the fraction that came from C is $\frac{1}{2n}$. So, if we account for attention, we now see that, as n gets large, it is in fact C that gets marginalized.

	actor		
Measure	C	X	Y
Betweenness	1	0	0
Flow	1	1	1
Random Walk	1	$\frac{8}{9}$	$\frac{2}{3}$
ABC-Centrality	$\frac{1}{2n-1}$	1	1

That is the implication of finite attention, and in a very high-level nutshell, that is the basic idea behind our proposed new measure of betweenness, which we call *Attentive Betweenness Centrality*, or ABC-centrality. Note that none of the other algorithms will have a dependence on n. In a world where C had bounded capacity/bandwidth/energy, C cannot possibly forward all traffic to all neighbors, and if C did, C would essentially become a spammer.

It is useful to explain exactly how the other measures failed to account for this finite attention. Betweenness simply asserts that C will forward to all nodes, and so given this infinite power, the quickest path is only through C. Flow betweenness asserts that A can tell C to send the information to B; not only is this rarely the case in practice, but A may not know the maximum flow path to B. In random walk betweenness, it is assumed that if C forwards the information to a random node like v, it will bounce back to C, and it will keep bouncing back to C until it reaches B; in reality if C sends the information

to v, that is likely the end of the story and the information is lost.

Information Flow. Before we develop any measure for betweenness, we had better return to the axioms. Betweenness is supposed to capture how much control an actor has over the *information flow of the network*. Well, in that case, we must first postulate the basic properties that information flow should have. We list the ones we consider fundamental.

I. *Forward Propagation.* An actor will not send information back along edges from where the information came.
II. *Locality.* An actor cannot process global information and perform global algorithms in determining how to forward information. An actor can only make use of its local neighborhood in deciding how to forward the information.
III. *Attention.* Actors have a finite attention they can give a piece of information. In the simplest case we can imagine an equal treatment of an actor's neighbors when deciding where to send a piece of information.
IV. *Multipath.* Information may flow along multiple paths to reach a destination, some longer than others. Controlling for attention, longer paths should be less valuable than shorter ones.

With respect to some notion of information flow satisfying these properties, the betweenness of a node v with respect to some other pair of nodes trying to exchange information should be the fraction of successful exchange that needs to pass through v.

Our Contributions. We present a simple betweenness measure, attentive betweenness centrality (ABC-centrality) that is based on a model of information flow that satisfies the basic properties above. It contains features of flow based methods, random walk based methods and incorporates preference for shortest paths. It is efficient to compute, having the same complexity as betweenness for unweighted graphs and better complexity for weighted graphs. A summary of the information flow models on which various betweenness measures are based is given below.

	Betweenness measure			
	Bet.	ABC	Flow	Rand Walk
Forw. Prop.	✓	✓	✓	✘
Locality	✘	✓	✘	✓
Attention	✘	✓	✘	✘
Non-shortest	✘	✓	✓	✓
Complexity	$O(mn)$	$O(mn)$	$O(m^2n)$	$O(mn^2)$

Bet=betweenness; ABC=our measure

Our algorithm contains a parameter $\alpha \in [0, 1]$ which determines the factor by which one prefers shorter paths over longer paths. With α closer to 0, our algorithm is an extension of betweenness that incorporates attention; With α closer 1, our algorithm starts to have features of degree centrality. Thus the general algorithm offers a spectrum of measures between these two, and a practitioner could pick the appropriate one.

We illustrate the benefits of our ABC-centrality on several stylized graphs. ABC-centrality has strong similarity to betweenness with the added benefit of attention and multipath; we demonstrate this not only on the stylized graphs but also on large scale applications to the IMDB actor-movie network and the DBLP author-paper academic network. Based on the expectation that high betweenness actors ought to be actors with a diversity of talents [11], we give a large scale quantitative study on IMDB that demonstrates that ABC-centrality is better able to capture this diversity than traditional betweenness.

In conclusion, our method is similar to betweenness with added benefits; it is just as efficient to compute; and, it performs better in practice and on stylized networks (at least within our limited experimental setup). As such, ABC-centrality ought to be considered as a serious alternative to betweenness when a betweenness measure of centrality is desired. We emphasize that our goal is not to present a centrality measure that dominates all other centrality measures, though it would be interesting to perform a large scale comparison of different types of centrality measures. Our goal is to address some of the deficiencies in the betweenness algorithms, which have been recognized in the past but not adequately addressed. Our hope is that the remainder of the paper will convince the reader that we do indeed address these deficiencies and provide a superior measure of betweenness.

Paper Organization. Next, we give the detailed description of the algorithm that computes ABC-centrality. We then give the comparison of various measures of centrality on the stylized networks as well as large scale validation on IMDB and DBLP. We conclude with a discussion of weighted graphs to which our algorithms seemlessly extend.

II. Algorithm for Computing ABC-Centrality

The ABC-centrality algorithm takes as input a value α and computes scores for each node. The high level description of the algorithm is as follows. The following small network in Figure 3 will be a useful concrete realization of the algorithm. For each node for example A, we imagine a unit of infor-

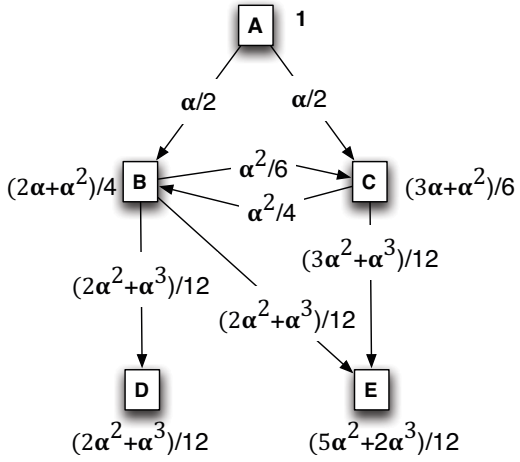


Fig. 3. Illustration of the computation of ABC-Centrality

mation being sent out. Every node that receives information propagates it out to its neighbors, dividing it equaly among its neighbors, with the exception that neighbors from where information came do not get any flow back. In this way, flow propagates from A through the network in a breadth-first-search manner and will eventually stop flowing after $O(m)$ steps. Now, every node (for example D) has received some flow. The fraction of this flow that passed through various other nodes on the way from A are what contribute to the centrality score of those nodes. So for example, for the flow from A to D, nodes B and C will be credited that fraction of the $(2\alpha^2 + \alpha^3)/12$ that reached D and passed through them. This entire process is repeated for information flow starting from every node, resulting in an $O(mn)$ final running time. As can be seen from Figure 3, the role α plays is to attenuate the information by the factor α for every edge the information traverses. So information that arrives to (for example) D via a longer path will get attenuated relative to information arriving via a shorter path.

To see that the algorithm satisfies the basic requirements of information flow, observe:

– *Forward propagation.* Information never flows back along links.

– *Locality.* A node forwards to all nodes who didn't already send it information uniformly.

– *Attention.* A node divides its effort among its neighbors (i.e. the information value gets split) as opposed to giving each neighbor 100% service, which could be unbounded for very large degree nodes.

– *Multipath.* Clearly information arrives to D using multiple paths of different lengths. The α parameter determines how much we prefer shorter paths. If $\alpha \to 0$ we only use shortest paths, and then the only difference between ABC and betweenness is attention.

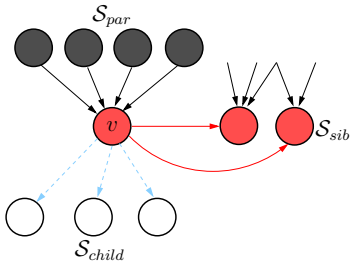
We now give the more detailed description of the algorithm, referring to Figure 3. The detailed pseudocode can be found in the appendix in Figure 10. The information flow model we will use results in an algorithm that is very similar to the algorithm for computing standard betweenness.

The algorithm is initiated at each node so for illustration, consider node A in Figure 3. The first step is the forward step in which information is propagated from A. The second step is the backward step in which we collect up the information that has flowed through each node.

Forward Step. Initiate one unit of flow from A. The forward propagation of flow proceeds in synchronous steps $t = 1, 2, \ldots$. We now describe the general process at a generic node v, and then illustrate the process with Figure 3. At every time step, either a node has received flow or not. The first time a node receives flow, it will propagate over the next two time steps. Suppose that v received information for the first time at step t; here is what happens at steps $t+1$ and $t+2$. We assume that v has access to information about the nodes in its neighborhood. In particular, it can categorize the nodes u in its neighborhood into three sets.

$$\begin{aligned} \mathcal{S}_{par}(v) &= \{u \text{ s.t. } u \text{ sent } v \text{ info.}\}; \\ \mathcal{S}_{sib}(v) &= \{u \text{ s.t. } u \text{ has info. but didn't send to } v\}; \\ \mathcal{S}_{child}(v) &= \{u \text{ s.t. } u \text{ has no information}\}. \end{aligned}$$

This information is all locally available, which is essential for the algorithm to be based on local information flow. The situation is illustrated below.



The grey nodes in $\mathcal{S}_{par}$ (your 'parents') have just forwarded information to v. Let x_t be the information coming from $\mathcal{S}_{par}$, and let $\delta = |\mathcal{S}_{sib}| + |\mathcal{S}_{child}|$. At step $t+1$, v forwards $\alpha x_t/\delta$ information to each neighbor in $\mathcal{S}_{sib}$ along the red arrows. This corresponds to dividing the information amongst its neighbors (finite attention) and first sending the information to your neighbors who already have information but didn't send it to you yet (your 'siblings'). You will receive information from these neighbors in this step as well. Let the amount of information you receive from your siblings be x_{t+1}. So your total information is now $x_t + x_{t+1}$. Node v now sends $\alpha(x_t + x_{t+1})/\delta$ to its 'children' nodes in $\mathcal{S}_{child}$ at time step $t+2$. After this point v will neither receive nor send any more information. One can show that this process is exactly analogous to a breadth-first (BFS) information propagation starting from A with the addition that information propagates during the even time steps with a BFS layer and during the odd time steps from BFS-layer ℓ to layer $\ell + 1$.

Every two time steps we process one layer in the BFS tree starting from the root A. So the first step to actually calculating the flow is to perform a BFS of your graph from A. To process a layer, information first flows within the layer. Then information flows from the layer to the next layer. Every (undirected) edge is processed exactly twice so the running time is $O(m)$. The process extends seemlessly if the graph is directed. For a weighted graph the only modification is that instead of uniformly splitting your information among your neighbors when you propagate, you split it in proportion to the weights. Similarly, the algorithm can be used with directed or undirected graphs.

We walk through the process in Figure 3. At step 1, A takes its unit and sends $\alpha/2$ to B, C. At step 2, B sends $\frac{\alpha}{3} \cdot \frac{\alpha}{2}$ to C; and C does something similar (B, C are siblings). Finally at step 3, B sends $\frac{\alpha}{3}$ of its total information to each of D, E; C sends $\frac{\alpha}{2}$ of its total information to E.

Backward Step. The backward step now allocates credit for all the flow to various nodes along the paths that the flow took, which we denote by c_v. We process all the nodes in

the BFS tree, layer by layer from bottom to top. The nodes at the lowest layer all get credit of 0. Then, each node v at the next level receives credit from nodes u at the level below and nodes at the same level. Let U be the set of all nodes v receives credit from and let x_u be the flow that reached u and let x_{vu} be that part of x_u which came from v. We then compute c_v as follows:

$$c_v \leftarrow \sum_{u \in U} (1 + c_u) \frac{x_{vu}}{x_u}$$

In other words, v is credited for that fraction of u's flow that came from v multiplied by the credit u has to give; v receives credit and also now has more credit to give. After processing all the children of nodes in the layer, the nodes process their siblings in a similar way, in batch mode (i.e. simultaneously). The process then moves up one layer. After the process is complete, the credit received by the nodes is the score. The entire process is repeated with a BFS starting from each node, and the scores are all averaged.

In our example in Figure 3, node B is responsible for some of the flow received at nodes C, D and E. It will get credit for each. For node C, the fraction

$$\frac{\alpha^2/6}{(3\alpha + \alpha^2)/6}$$

is sent by B, so when B processes C, it will get that fraction of the credit that C has to give. Similarly, B gets all the credit for the flow received by D (and since D has 1 to give, this will be 1 received by B – B will give some of its total credit to C and the rest to A); finally, D will also receive a fraction

$$\frac{(2\alpha^2 + \alpha^3)/12}{(5\alpha^2 + 2\alpha^3)/12}$$

of the credit E has to give (which is also 1).

Again, the parameter α, from this simple example of Figure 3, can be seen to determine how important longer paths are. Our measure becomes closer to betweenness as α approaches 0. Second, the more such paths the node is on, the better, however, attention plays a role in determining how much credit the node gets if alternative high attention paths are available. The flow is only forward going unlike random walk based diffusion.

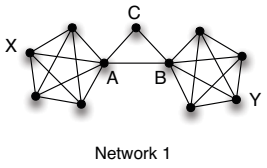
III. Comparison of Centrality Methods

We now compare ABC-Centrality with various other centrality measures in different networks. We focus on unweighted networks, because betweenness is predominantly studied in unweighted networks. However, our measure seamlessly extends to weighted graphs and we give a brief discussion of this toward the end of the paper. Table I lists the algorithms studied in this paper. Note that, we are able to study FLOW and RW only on very small networks because their computational complexity does not allow them to scale to large networks. While PG, DEG and CL are generally different than betweenness, we are including them to show similarities and differences between these algorithms for different data sets.

Abbr	Description
ABC^{α}	ABC-Centrality we report on $\alpha = 0.001, 0.5, 1$
BET	Betweenness
ABC^{0^+}	ABC-Centrality, $\alpha = 0.001$
$ABC^{0.5}$	ABC-Centrality, $\alpha = 0.5$
ABC^{1}	ABC-Centrality, $\alpha = 1$
FLOW	Flow Betweenness [6]
RW	Random Walk Betweenness [10]
DEG	Degree Centrality
CL	Closeness Centrality
PG	Pagerank (using 0.85) [12]

TABLE I
CENTRALITY MEASURES STUDIED IN THIS PAPER

a) Stylized Networks: We begin with stylized networks that highlight the similarities and differences between various centrality measures. The first two networks are taken from [10], and the third one is from Figure 1. The scores of different algorithms are scaled to a comparable range.



Network 1

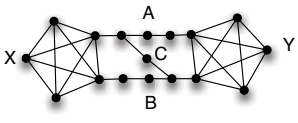
Measure	A, B	C	X, Y
BET	1.00	0.00	0.00
FLOW	1.00	0.45	0.10
RW	1.00	0.49	0.40
PG	1.00	0.39	0.69
ABC^{1}	1.00	0.51	0.19
$ABC^{0.5}$	1.00	0.33	0.12
ABC^{0^+}	1.00	0^+	0^+

0^+ is number slightly above zero

Fig. 4. Example network from [10].

In Figure 4[1], nodes A and B have the highest betweenness as they lie on all shortest paths between the left and the right cliques. For these paths, node C gets no credit as the path is slightly longer if we travelled through node C. All the other algorithms allow non-shortest path flow, and so assign some value to C. The surprising thing about PG and RW is that they assign surprisingly high scores to X, Y even though these nodes are not very critical to the flow of information in this network.

For the previous network, all the multipath algorithms found some value for the node C. The next network in Figure 5 shows how flow-betweenness can ignore useful nodes because max-flow paths tend to be non-overlapping. Now FLOW completely ignores C, even though C can play an important role in



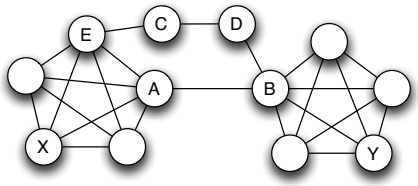
Network 2

Measure	A, B	C	X, Y
BET	1.00	0.81	0.00
FLOW	1.00	0^+	0.06
RW	1.00	0.84	0.59
PG	0.68	0.68	1.00
ABC^{1}	1.00	0.75	0.31
$ABC^{0.5}$	1.00	0.75	0.19
ABC^{0^+}	1.00	0.66	0^+

Fig. 5. Example network from [10].

[1]Scores in each row are rescaled so that the maximum is 1 for Figure 4, 5, and 6.

information flow. However PG and RW, again, attribute high score to X, Y which are peripheral nodes. ABC-Centrality seems to do a good job on all fronts for α chosen as some reasonable number like 0.5. Our final stylized network is the one we used in the introduction, reproduced here in Figure 6 for convenience. A superficial analysis of this network might



Measure	A	B	C	D	E	X	Y
BET	0.71	1.00	0.09	0.16	0.18	0.00	0.00
FLOW	0.62	1.00	0.46	0.46	0.62	0.08	0.08
RW	0.75	1.00	0.48	0.50	0.60	0.50	0.38
PG	0.83	1.00	0.40	0.40	0.83	0.66	0.66
ABC^1	0.62	1.00	0.27	0.30	0.35	0.16	0.16
ABC$^{0.5}$	0.63	1.00	0.24	0.27	0.29	0.10	0.10
ABC$^{0^+}$	0.75	1.00	0.12	0.17	0.29	0^+	0^+

Fig. 6. Example network from the introduction.

go as follows. B is most critical because without B, there is no information flow between the left and the right. A is not as important as in the first network because there are now paths available through E, but the paths through E are slightly longer. All shortest paths between the left and right that pass through C or D must use E, so one might expect E, C, D to be roughly equally critical; however E should dominate a little since there are just more information flow paths available that use E than C, D. X, Y are as usual peripheral nodes, and should only be marginally critical. So we expect

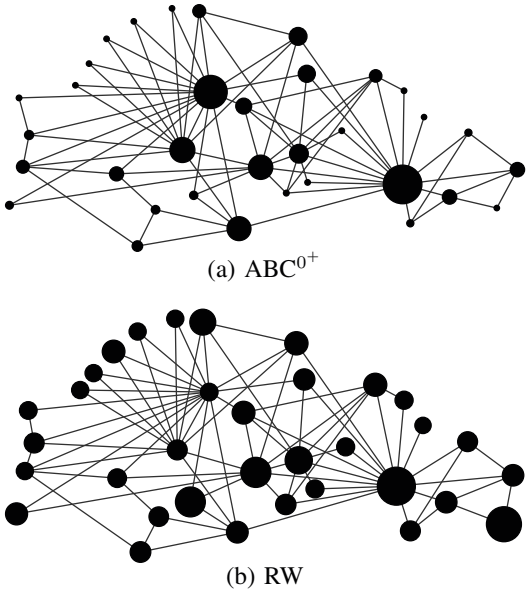
$$B > A > E \gtrapprox C \approx D \gg X \approx Y.$$

An examination of the scores from all the measures reveals that the ABC-Centrality measures are the ones that most closely realize this expectation.

Conclusions from Stylized Graphs First we observe that ABC-Centrality and betweenness are different. When $\alpha = 0$ the two are similar in spirit, focussing on shortest paths, but ABC-Centrality also takes into account attention. When α is large, ABC-Centrality starts to emphasize the shortest path less and now focuses more on diversity of paths, and attention.

On these stylized networks, ABC-Centrality seems to deliver the results that you would expect of a measure of criticality: peripheral nodes have low criticality; nodes that are on short paths are favored, but if they are 'overloaded', they will become disfavored with respect to nodes on longer paths that are not as overloaded.

So we see that ABC-Centrality accounts for some of the shortcomings of betweenness. Does it still retain the essence of betweenness? In other words, we may have solved the 'problems' with betweenness, but in doing so we may have constructed a measure that is completely different. As we will soon see, this is *not* the case. We have retained the essence of betweenness, while improving along the dimensions that betweenness is lacking.



(a) ABC$^{0^+}$

(b) RW

Fig. 7. A comprison of ABC & RW centrality scores for the Karate Club Network (the area of the node is proportional to the score). (a) ABC does a good job of highlighting the critical nodes for information flow. ABC has essentially a correlation of 1 with betweeness and so the scores for ABC are essentially the same scores obtained by BET. (b) RW does not differentiate between nodes very effectively and there are some unusually high scores for outlier nodes (such as the bottom right node); RW also does not seem to capture the high centrality of some intuitively critical nodes in the left cluster.

b) Karate Club Network: We will now look at the various centrality measures on a real social network of moderate size. The goal is to compare how well these different measures capture betweenness. We use the Zachary Karate Club Network [13] which contains 34 nodes and 78 undirected edges representing friendship relationships.

The similarity between two measures of centrality can be measured by the correlation of the centrality scores. The correlations between the scores of different algorithms is given in Table II.

For this Karate-club network, random walk (RW) is an outlier. The centrality scores are illustrated in Figure 7, which shows the results for ABC (essentially the same as BET) and RW. From the figure, we can observe that ABC picks up the intuitively critical nodes; RW does not differentiate significantly between the nodes, produces some unusually high scores for outlier nodes (for example the bottom right node) and unusually low scores for seemingly critical nodes in the centers of clusters. The results of RW do not conform to one's expectation given that the Karate Club network was formed from the fracture of a single group into two groups with two clear leaders in each group. From the table, it is clear that ABC is the most correlated with BET (among the algorithms we tested) and the correlation increases as α goes down, as expected. Both degree and flow are the other two centrality

	BET	ABC^1	ABC$^{0^+}$	RW	DEG	CL	FLOW	PG
BET	1	**.98**	**1**$^-$	.51	.92	.72	.95	.92
ABC^1	.98	1	.98	.51	.96	.77	.96	.97
ABC$^{0^+}$	1$^-$	.98	1	.51	.92	.73	.96	.93
RW	.51	.51	.51	1	.41	.32	.53	.42
DEG	.92	.96	.92	.41	1	.77	.91	1$^-$
CL	.72	.77	.73	.32	.77	1	.59	.74
FLOW	.95	.96	.96	.53	.91	.59	1	.93
PG	.92	.97	.93	.42	1$^-$	.74	.93	1

TABLE II
CORRELATION OF VARIOUS CENTRALITY SCORES FOR THE ZACHARY KARATE CLUB. FOR TWO MEASURES WHICH ASSIGN SCORES $s_1, \ldots, s_n$ AND $t_1, \ldots, t_n$ TO THE NODES IN THE NETWORK, THE CORRELATION IS DEFINED AS

$$\rho = \frac{\sum_{i=1}^{n}(s_i - \bar{s})(t_i - \bar{t})}{\sigma_s \sigma_t},$$

WHERE $\bar{s}, \bar{t}$ ARE THE AVERAGE VALUES AND σ_s^2, σ_t^2 ARE THE VARIANCES.

measures highly correlated with betweenness on this small network.

Conclusion. For this network ABC-Centrality captures the essence of betweenness, more so than all the other measures, and this conclusion is robust to the choice of α in our algorithm. Since the Karate-club network is a typical bi-partisan network, we expect this conclusion to generalize to other social networks.

So, at the small scale network, everything seems rosy. Let's now take the ball game to an entirely different playing field – very large scale social networks.

c) Internet Movie Database (IMDB): We also study a subset of the Internet Movie Database [2] which consists of actors and directors who featured in movies from 2000 to 2009. For movies, we only took the top 3 actors in the movie given by the order of appearance. This eliminates actors with small and cameo appearances in the movies. The resulting graph has 32,557 nodes, each node is either an actor, actress or a director. Two nodes are linked if they participated in the same movie. The unweighted network has 82,832 edges. One could also weight the edge by the number of movies two nodes have in common to obtain a weighted graph, but for now we use the unweighted version. We will briefly discuss the weighted versions later. We show the correlations between the different centrality measures in Table III.

	BET	ABC^1	ABC$^{0^+}$	DEG	CL	PG
BET	1	**.95**	**1**$^-$	.72	.27	.75
ABC^1	.95	1	.97	.76	.38	.78
ABC$^{0^+}$	.99	.97	1	.73	.29	.76
DEG	.72	.76	.73	1	.41	.94
CL	.27	.38	.29	.41	1	.27
PG	.75	.78	.76	.94	.27	1

TABLE III
CORRELATION OF VARIOUS CENTRALITY SCORES FOR THE UNWEIGHTED IMDB NETWORK OF 2000S

First, observe that we have not reported results for RW and

[2]imdb.com

	BET	ABC^1	ABC$^{0^+}$	DEG	CL	PG
BET	1	**.90**	**.97**	.80	.41	.83
ABC^1	.90	1	.97	.87	.58	.93
ABC$^{0^+}$	.97	.97	1	.84	.49	.88
DEG	.80	.87	.84	1	.66	.97
CL	.41	.58	.49	.66	1	.60
PG	.83	.93	.88	.97	.60	1

TABLE IV
CORRELATION OF VARIOUS CENTRALITY SCORES FOR THE UNWEIGHTED DBLP NETWORK

FLOW. This is due to their inability to scale to such large networks. Among the remaining algorithms, the conclusion is similar. ABC-Centrality is again the measure that is by far the most correlated with betweenness. Now, even degree is not very correlated with betweenness. We see that pagerank and degree are generally highly correlated with each other as it was observed in previous literature [14]. We also show the scatter plot of the ABC-Centrality measure versus betweenness for two different α values in Figure 8. As seen in this figure, the relationship is near-linear and the distribution of the values does not change drastically as a function of α.

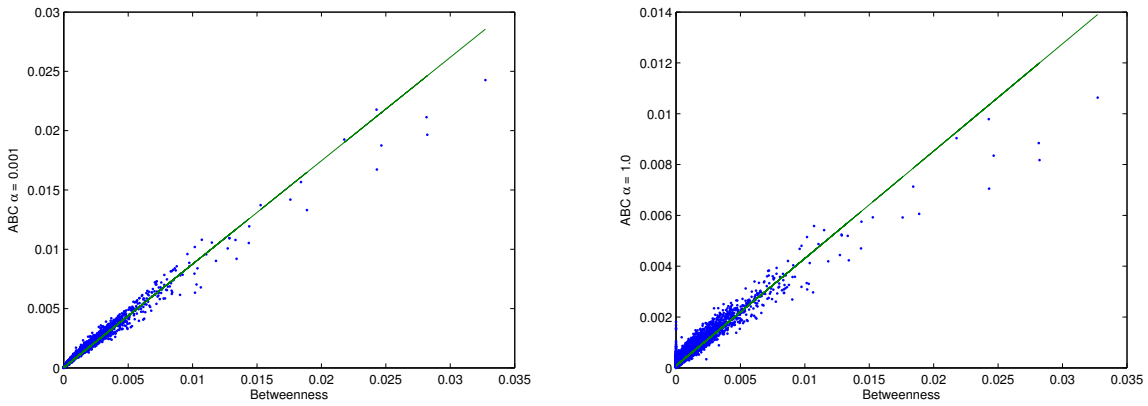
Conclusion. For a large scale social network, the same result holds. ABC-Centrality is the measure that most accurately captures the essence of betweenness.

d) Academic Collaboration Network (DBLP): Finally, we study the DBLP network [3] containing researchers in Computer Science and the papers that they wrote. In this network, two authors are connected if they co-wrote a paper. We choose only those authors with more than 6 papers. The DBLP graph we study has 74,443 authors and 417,397 edges. The correlation between different centrality measures is given in Table IV.

Again, ABC-Centrality is the most correlated measure to betweenness in DBLP as with all our other networks, and this is independent of the value of α. However, there is a more significant variability as we change α for this network. We attribute this to the higher number of edges in this network. In fact, the degree distribution in this network is heavy tailed in that most authors have few co-authors (2-3) but a small number of authors have an unusually high number of co-authors. For example, according to Microsoft Academic Search [4], Thomas S. Huang has 745 co-authors and Alberto L. Sangiovanni-Vincentelli has 752 co-authors. There are also data curation problems in DBLP, especially due to error in entity resolution in shorter and more common names leading to multiple authors being merged into a single entity. In addition, it seems that in this network nodes with high degree tend to also have high pagerank and high betweenness. This means that authors that tend to bridge multiple communities tend to collaboratorate with many others. As a result, the high degree nodes dominate all the various centrality measures. Nevertheless, we still end up with a measure that has captured

[3]dblp.uni-trier.de

[4]academic.research.microsoft.com



(a) $\alpha = 0.001$ (b) $\alpha = 1$

Fig. 8. Scatter plot showing the correlation of the ABC centrality values vs. betweenness values in IMDB for different α values. The points are nearly on a line with extremely high betweeness values tending to be lower for ABC (this is the impact of attention) and the low betweeness values tending to be higher (this is the impact of multi-path).

the essence of betweenness.

IV. External Validation

We have an interesting situation. We have shown that ABC-Centrality is different than betweenness on stylized networks; but, different in the right way, capturing elements of criticality that betweenness failed to capture. But, on large networks, we see that ABC-Centrality and betweenness have a correlation of almost 1. So, if they are so similar on large real networks, is there any reason to use one over the other. We now show quantitatively in a large scale study that even on these large networks, though the correlation is nearly 1, the small differences lead to consequences. In particular, using an external charasteristic, we demonstrate a quantifiable improvement that ABC-Centrality has over betweenness.

Diversity. It is believed that betweenness is a measure of an actor's diversity [11]. While betweenness is not the sole measure of diversity, it is one of the important indicators. In other words, a node with large betweenness means that the node will typically belong to multiple communities. But you must have diverse talents if you can belong to different communities.

Take actors. An actor belonging to different actor communities (as defined by the movies they act in) probably means they have the ability to play multiple different types of roles. The actors of highest betweenness in IMDB are

Rank	Actor
1	Michael Masden
2	David Carradine
3	James Russo
4	Joe Estevez
5	Eric Roberts
⋮	⋮

You will observe that these are not high profile actors. They are multi-faceted actors who will take on secondary roles in multiple different types of movies. So for example Michael Masden is a 'generic villian' who might appear in a romance, action, adventure, thriller, etc.

For an actor, we can use the various different types of movies they act in to quantitatively compute an index of diversity. Suppose there are genres $g_1, \ldots, g_k$, and consider an actor who acted in three movies with the genres

Movie	Genres
1	g_1
2	g_1, g_2
3	g_1, g_3

We compute the fraction of that actor's effort spent on each genre by treating each movie to be 1 unit of effort and then splitting that effort among the genres of the movie. So in this particular example, the actor's efforts for each genre are

$$g_1: \ 1+\tfrac{1}{2}+\tfrac{1}{2}=2; \ g_2: \ 0+\tfrac{1}{2}+0=\tfrac{1}{2}; \ g_2: \ 0+0+\tfrac{1}{2}=\tfrac{1}{2}.$$

Normalizing these efforts to sum to 1, we get a probability distribution that reflects the fraction of an actor's effort spent on each genre. In this case the probability distribution would be $(\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$. This genre probability distribution represents how multi-faceted or diverse the actor is. We can measure this diversity quantitatively using the entropy of the probability distribution, a well known measure of how concentrated the distribution is. Let $\boldsymbol{p} = (p_1, \ldots, p_k)$ be the genre probability distribution for an actor. We define its diversity D by

$$D = H(\boldsymbol{p}) = -\sum_{i=1}^{k} p_i \log p_i.$$

Note that $0 \log 0 = 0$ and $0 \leq D \leq \log k$.

We may now quantitatively ask: "How well does the ABC-Centrality of an actor track its diversity D, as compared to betweenness?" More specifically, we can ask how well ABC-Centrality correlates with diversity D as compared to the

correlation of betweenness with D. As a function of the degree of a node, we plot the *increased* correlation with diversity that is offered by ABC-Centrality in Figure 9.

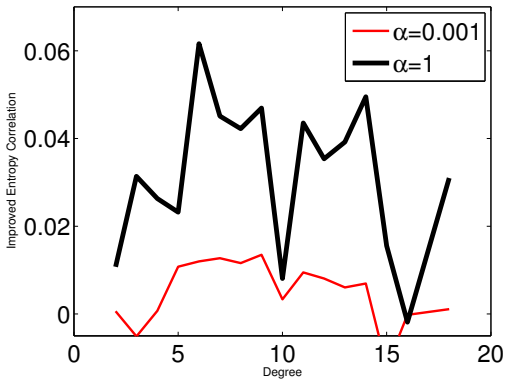


Fig. 9. Improvement of the correlation with an actors diversity index D that ABC-Centrality offers over betweenness.

As we can observe, the correlation coefficient can increase by as much as 0.06 for a large α. For a small alpha, ABC-Centrality is very similar to betweenness except with the addition of attention. Now, the improvement is less, but still present on average. This means that attention alone seems to capture something more than just pure betweenness.

V. Weighted Graphs

ABC-Centrality can be easily extended to weighted graphs. In fact, the pseudocode in the appendix is described for weighted graphs. As a closing experiment, we show the correlations between the various measures for our large scale networks treated as weighted graphs. Since the weight is meant to represent social distance, low weights represent close relationships. The weights in IMDB are given by a variation of the Adamic-Adar measure [15]: given a pair of actors, we find all the movies m these actors have in common and sum $1/\log z$ over all such movies where z is the number of actors in movie m. The distance is given by the reciprocal of this value. Similarly in DBLP, the number of papers that authors have in common are used by taking into account the total number of authors in each paper.

The score correlations for IMDB are given in Table V and for DBLP in Table VI. We note that in IMDB, our algorithm continues to be very correlated to betweenness.

However, for DBLP, none of the other algorithms are very similar to betweenness. In fact, we see that ABC is highly correlated with degree and pagerank. This may be related to the presence of extremely high degree nodes which swamp the other measures. Nevertheless, ABC-Centrality is still (essentially) the most correlated with betweenness (together with PG).

VI. Conclusion

Our goal was to present a measure of betweenness that improves upon the traditional measure. We broadly defined betweenness as the control a node has over the information flow in a network. This lead to the need to formulate the basic principles of information flow. Given a model for information flow that satisfies these basic principles, the betweenness is defined in the standard manner. Take two nodes and send information between them. The amount of that information that needs to pass through a generic third node v is the betweenness of v for this pair. Now average over all possible pairs.

Our principles of information flow are very general and so this opens the door to a family of betweenness measures. We used perhaps the simplest model of information flow that is forward propagating, local, attention-sensitive and multipath. Even still, the results show that we have captured the essence of the traditional betweenness measure; we overcome its difficulties; and we demonstrate quantifiable improved performance on real, large-scale social networks.

Our approach to quantifying the quality of a betweenness measure appears to be novel, and is just a begining. We hope to extend our study within this framework of external validation to a number of different measures. Since centrality in general, and betweenness in particular are considered important measures of prominence, such approaches to validation on large scale networks could be valuable.

	BET	ABC^1	ABC^{0^+}	DEG	CL	PG
BET	1	**.92**	**.96**	.70	.27	.73
ABC^1	.92	1	.97	.76	.38	.77
ABC^{0^+}	.96	.97	1	.73	.30	.76
DEG	.70	.76	.73	1	.40	.92
CL	.27	.38	.30	.40	1	.26
PG	.73	.77	.76	.92	.26	1

TABLE V
Correlation of various centrality scores for the weighted IMDB Network of 2000s

	BET	ABC^1	ABC^{0^+}	DEG	CL	PG
BET	1	**.64**	**.64**	.56	.28	**.65**
ABC^1	.64	1	.97	.85	.40	.84
ABC^{0^+}	.64	.97	1	.83	.31	.78
DEG	.56	.85	.83	1	.45	.83
CL	.28	.40	.31	.45	1	.40
PG	.65	.84	.78	.83	.40	1

TABLE VI
Correlation of various centrality scores for the weighted DBLP Network

Acknowledgment

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] K. Faust and S. Wasserman, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[2] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, Mar. 1977.

[3] ——, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[4] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, pp. 163–177, 2001.

[5] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Social Networks*, vol. 11, no. 1, pp. 1–37, Mar. 1989.

[6] L. C. Freeman, S. P. Borgatti, and D. R. White, "Centrality in valued graphs: A measure of betweenness based on network flow," *Social Networks*, vol. 13, no. 2, pp. 141–154, 1991.

[7] S. Milgram, "The Small World Problem," *Psychology Today*, vol. 2, pp. 60–67, 1967.

[8] P. S. Dodds, R. Muhamad, and D. J. Watts, "An Experimental Study of Search in Global Social Networks," *Science*, vol. 301, no. 5634, pp. 827–829, Aug. 2003.

[9] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, 1st ed. Prentice Hall, Feb. 1993.

[10] M. Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. 27, no. 1, pp. 39–54, 2005.

[11] L. Leydesdorff and I. Rafols, "Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations," *Journal of Informetrics*, vol. 5, no. 1, pp. 87 – 100, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1751157710000854

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the ACM WWW Conference*, 1998, pp. 107–117.

[13] Z. W., "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, pp. 452–473, 1977.

[14] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas, "Link analysis ranking: Algorithms, theory, and experiments," *ACM Transactions on Internet Technology*, vol. 5, no. 1, pp. 231–297, 2005.

[15] L. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

APPENDIX

We use the notation $\Gamma(u,i)$ to denote the neighbors of node u at level i and $norm(u,i) = \sum_{v \in (\Gamma(u,i) \cup \Gamma(u,i))} weight(u,v)$ for the sum of the forward weights. The $norm$ function disregards the weight of the edges at level i-1, and considers only the edges at level i or i+1.

---

function SINGLESOURCEABC(Graph g, s, α)
  V $\leftarrow$ vertices(g);
  $pflow(v) = flow(v) = 0$ for all $v \in V$
  $pflow(s) = flow(s) = 1$; n = max_bfs_level(g)
  for i: 1 to n **do**
    for v is a node in V at bfs level i **do**
      $flow(v) = \sum_{u \in \Gamma(v,i-1)} \frac{\alpha.pflow(u).weight(u,v)}{norm(u,i)}$
      $pflow(v) = flow(v)$
      $flow(v) += \sum_{u \in \Gamma(v,i)} \frac{\alpha.flow(u).weight(u,v)}{norm(u,i)}$
    end for
  end for
  $score(v) = 0$ for all v
  for i: n to 1 **do**
    for v is a node in V at bfs level i **do**
      $score(v) = \sum_u \frac{(1+score(u)).(\alpha.flow(v).weight(v,u))}{flow(u).norm(v,i)}$
      where $u \in \Gamma(v, i+1)$
      $score(v) += \sum_u \frac{(1+score(u)).(\alpha.pflow(v).weight(v,u))}{flow(u).norm(v,i)}$
      where $u \in \Gamma(v,i)$
    end for
  end for
  return $score$
end function

function ABC-CENTRALITY(Graph g, α)
  V $\leftarrow$ vertices(g); ABCScores(v)=0 for all nodes in V.
  for all nodes v in V **do**
    ABCScores += SingleSourceABC(g,v,α)
      ▷ add scores for each vertex
  end for
  for all nodes v in V **do**
    ABCScores(v) = ABCScores(v)/$|V|$
  end for
  return ABCScores
end function

Fig. 10. The ABC-centrality algorithm