Process-driven betweenness centrality measures

Mareike Bockholt[1] and Katharina A. Zweig[1]

Department of Computer Science, University of Kaiserslautern
Gottlieb-Daimler-Straße 48, 67663 Kaiserslautern, Germany
{mareike.bockholt,zweig}@cs.uni-kl.de

Abstract. In network analysis, it is often desired to determine the most central node of a network, for example for identifying the most influential individual in a social network. Borgatti states that almost all centrality measures assume that there exists a process moving through the network from node to node [4]. A node is then considered as central if it is important with respect to the underlying process. One often used measure is the betweenness centrality which is supposed to measure to which extent a node is "between" all other nodes by counting on how many shortest paths a node lies. However, most centrality indices make implicit assumptions about the underlying process. However, data containing a network and trajectories that a process takes on this network, are available: this can be used for computing the centrality. Hence, in this work, we use existing data sets, human paths through the Wikipedia network, human solutions of a game in the game's state space, and passengers' travels between US American airports, in order to (i) test the assumptions of the betweenness centrality for these processes, and (ii) derive several variants of a "process-driven betweenness centrality" using information about the network process. The comparison of the resulting node rankings yields that there are nodes which are stable with respect to their ranking while others in- or decrease in importance dramatically.

Keywords: network analysis, centrality measures, network processes, path data analysis

1 Introduction

An often performed task in network analysis is the identification of the most important nodes in a network. The goal might be to find the most influential individual in a social network, the most vulnerable location in a transportation network, or the leader in a terrorist network [6,17]. The identification of such nodes is usually done with a centrality measure which computes a value for each node of the network based on the network structure [2,20,10]. The concept of centrality in networks was first introduced by Bavelas in the late 1940s who considered human communication networks [2]. Inspired by this idea, a large number of different methods for measuring the centrality of a node were proposed in the following decades, where the best known centrality measures are degree centrality [9], closeness centrality [9], betweenness centrality [9,1], and Eigenvector centrality [3] (for an overview, cf. [5] or [14]).

An important contribution was made by Borgatti who states that almost all centrality indices are based on the assumption that there is some kind of traffic (or communication or process) flowing through the network by moving from node to node [4]. This might be the propagation of information in a social network, packages being routed through the WWW, or the spreading of a disease in human interaction networks. A node is then considered as central, i.e., is assigned a high value of the measure, if it is somehow important with respect to this underlying process. However, different centrality measures make different assumptions about the process flowing through the network. Some measures are based on shortest paths in the network, assuming that the underlying process moves on shortest paths. Other assume that, if whatever flows through the network, is at one node will spread simultaneously to all the node's neighbors, while others assume that it can only be at one node at a time.

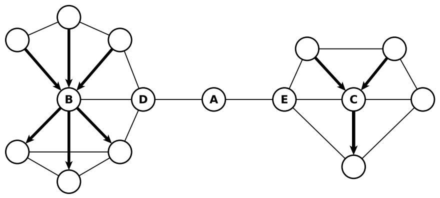


Fig. 1. Introductory example. Node A can be seen as a gatekeeper between the two node groups. Classic betweenness centrality hence assigns a high value to A. If, however, the network process only moves within the two groups (indicated by boldly drawn edges), should A still be considered as the most central node of the network?

Borgatti provides a typology of the most popular centrality measures by considering the network process [4] and argues that centrality measure values for a network can only be interpreted in a meaningful way, if the assumptions of the measure respect the properties of the process. He identifies two different dimensions by which the process can vary: First, on which type of trajectory does the process move through the network, and second, how does the process spread from node to node? For the first dimension, he differentiates between shortest paths, paths (not necessarily shortest, but nodes and edges can only occur at most once in it), trails (edges might occur several times in it, while nodes cannot be repeated), and walks (in which nodes and edges might occur several times). The second identified dimension is the "mechanism of node-to-node transmission": a process might *transfer* a good from node to node (e.g., a package), or it passes something to the next node by *duplicating* it: whatever flows through the network is passed to the next node and simultaneously stays at the current node (e.g., information flowing through a network or a viral infection: it will still be at the current node after it was passed to another node). Duplication can take place in a serial (one neighbor at once) or in a parallel manner (simultaneously to all neighbors). This categorization is a helpful tool for choosing an appropriate centrality measure, given a network and a process on the network.

The betweenness centrality is an often used centrality measure in social network analysis [9,1]. The idea is to measure to which extent a node v is positioned "between" all other nodes. A node with a high betweenness value is assumed to have control over other nodes: by passing information or withholding it, it can control the information flow. Consider for example the graph in Fig. 1 in which node A can be seen as gatekeeper between the left subgraph and the right subgraph: A can prevent information being passed to the other subgraph because all paths between the two subgraphs pass through A. Betweenness centrality counts how many shortest paths from any node s to any other node t pass through the given node v and averages over all node pairs (s, t). In order to account for cases in which there are many shortest paths between s and t, but only a few of them pass through v, the measure normalizes by the total number of shortest paths between s and t. Formally, with σ_{st} denoting the number of shortest paths from a node s to a node t (with $\sigma_{st} = 1$ if $s = t$) and with $\sigma_{st}(v)$ denoting the number of shortest paths from s to t that pass through v,we can define the betweenness centrality for a node v as[1]

$$B_c(v) = \sum_{\substack{s \in V, \\ s \neq v}} \sum_{\substack{t \in V, \\ s \neq t \neq v}} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

According to Borgatti [4], this measure is appropriate for networks on which the process of interest literally *moves* from node to node by a transfer mechanism and travels on shortest paths. This becomes obvious in Fig. 1: if there was added two additional nodes D' and E', and edges linking D to D', D' to E' , and E' to E, the high importance of A as a gatekeeper can only be still justified if the process exclusively takes shortest paths and can only take *one* path at once.

When taking a closer at the formula, it is easy to see that there are even more simplifying assumptions in this measure: it assumed that the process flows between any node pair (s, t) and the amount of traffic flowing from s to t is equal (and equally important) for all node pairs. For a given network with a process fulfilling the two main assumptions, these further assumptions are justifiable: if there is no other information about the process available, these assumptions are the best possible. However, if the information of how the process moves through the network is available, the importance of the nodes can be measured differently. Consider again Fig. 1 in which the bold directed edges indicate that there is communication between those nodes and no communication between the other nodes. Hence, in both groups, there are many entities moving from one of the upper nodes to (the) one node below, using the nodes B and C, respectively, but no entity moving from one group to the other. Although the assumptions of moving on shortest paths and by a transfer mechanism are met, should the node A still be considered as the most central node if there is actually *no* communication at all between those two groups?

However, there are data sets available containing a network structure and trajectories that the network's process has taken. It is therefore possible to actually use the information of how the process moves through the network in order to assign a centrality value to the nodes. It is clear that this is a different approach than the classic betweenness centrality or other existing centrality measures: classic centrality indices solely use the *structure* of the given graph in

[1] In literature, it is often noted as C^B. Since we are only considering this centrality measure and for a better readability in Sec. 6, we use this notation.

order to compute centrality values for the nodes – the choice which centrality index is appropriate for this network can then taken based on the knowledge about the properties of the process. Here, the information about the process is already used for *computing* the centrality value.

Hence, the contribution of this work is the following: we use three data sets in order to (i) investigate whether the assumptions of the betweenness centrality are met by those processes, and (ii) incorporate the information about the process contained in the data sets into a *process-driven betweenness centrality* (PDBC). Several variants of PDBC measures are introduced in order to analyse which piece of information affects the node ranking in which manner. For example, one variant only counts the shortest paths between those node pairs which are source and destination of the process at least once. Another variant does not count the number of shortest paths, but counts the number of actually used trajectories in which a node is contained. Those variants are then applied on the available data sets and the resulting node rankings of each variant compared to each other. It can be observed that the resulting rankings show a high correlation to each other, but there are nodes whose importance in- or decreases considerably.

This article is therefore structured as follows: Section 2 introduces the necessary definitions and notations before Sec. 3 presents related work in this area. Section 4 gives a description of the available data sets. Section 5 discusses the assumptions of the betweenness centrality and tests whether those assumptions are met in the data sets. Section 6 will introduce four variants of a process-driven betweenness centrality and Sec. 7 discusses the results of the PDBC measures on the given data sets, before Sec. 8 summarises the articles and gives an outlook for future work.

2 Definitions

$G = (V, E)$ denotes a directed simple unweighted graph with vertex set V, and edge set $E \subseteq V \times V$. A *path* is an alternating (finite) sequence of nodes and edges, $P = (v_1, e_1, v_2, \ldots, e_{k-1}, v_k)$ with $v_i \in V$ and $e_j = (v_j, v_{j+1}) \in E$ for all $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, k-1\}$, respectively[2]. Since G is simple, P is uniquely determined by its node sequence and the notation can be simplified to $P = (v_1, v_2, \ldots, v_k)$. The length of a path P ($|P|$) is defined as its number of (not necessarily distinct) edges. The start node of the path P is denoted by $s(P) = v_0$, the end node by $t(P) = v_k$ (occasionally referred to as source/start and destination/target). Let $d(v, w)$ denote the length of the shortest path from node v to node w. If w cannot be reached from v, we set $d(v, w) := \infty$.

3 Related Work

As illustrated in Sec. 1, the betweennness centrality contains several assumptions. Since these assumptions are not met in all networks and all processes, there have

[2] Note that we do not require the nodes and edges to be pairwise distinct. In some literature, P would be referred to as a walk.

been proposed many variants of the betweenness centrality. The most prominent variants questioning the assumption of shortest paths are the flow betweenness centrality of Freeman [11], and the random walk betweenness centrality by Newman [16], but there are also variants as routing betweenness [7] or variants for dynamic networks where paths up to a certain factor longer than the shortest one contribute to the centrality [6]. A variant questioning the assumption that all paths contribute equally to centrality value is the length-scaled betweenness [5].

Path analysis The idea of analysing the process moving through a network is not new. In many real-world networks, entities use the network to navigate through it by moving from one node to the other. This includes Internet users surfing the web yielding clickstream data, or passengers travelling through a transportation network. The navigation of an entity in a network to a target node, but with only local information about the network structure is called *decentralized search.* The fact that humans are often able to find surprisingly short paths through a network was already illustrated by Milgram in his famous small world experiment in 1967 [15]. An answer of *how* humans actually find these short paths was not known before Kleinberg investigated which effect the network structure has on the performance of any decentralized search algorithm [13]. West and Leskovec analysed the navigation of information seekers in the Wikipedia networks [23] and also find that human paths are surprisingly short, and show similar characteristics in their structure. This is also found by Iyengar et al. [21] who considered human navigation in a word game.

Combination of centrality and path analysis There are approaches of using the information about the network process in order to infer knowledge about the network itself: West et al. [24] use the paths of information seekers in the Wikipedia network to compute a semantic similarity of the contained articles. Rosvall et al. [19] show how real-word pathway data incorporated into a second-order Markov model has an effect on the node rankings in an approach generalising PageRank, and can derive communities of the network by the network's usage patterns. Zheng [26] identify popular places based on GPS sequences of travellers. Also based on GPS trajectories, the approach of Yuan et al. [25] makes use of Taxi drivers' trajectories in order to compute the effectively quickest path between two places in a city. Dorn et al. [8] can show that the results of betweenness centrality in the air transportation network significantly change if the centrality considers the number of actually taken paths traversing through an airport instead of all possible shortest paths.

4 Data sets

The goal is to investigate whether the assumptions built into the betweenness centrality are actually met by a data set containing the information how a process moves in a network. Since we consider the betweenness centrality, it is, according to Borgatti [4], only meaningful to consider processes which (i) move by a transfer mechanism, and (ii) move through the network with a target, i.e., a predetermined node to reach. The first condition implies that the movement of the entity can be modelled as a path. Data sets appropriate to use in this work hence need to fulfil the following requirements

(i) it contains a network structure, given as graph $G = (V, E)$

(ii) it contains a set of paths of one or several entities moving through the network, given as $\mathcal{P} = \{P_1, \dots, P_\ell\}$

(iii) the entities move through the network with a predetermined goal to reach,

(iv) that they aim to reach as soon as possible.

We use the following data sets (see also Tab. 1).

Wikispeedia This data set provided by West et al. [23,24] contains (a subset of) the network of Wikipedia articles where a node represents an article and there exists an edge from one node to another if there exists a link in the one article leading to the other article. The paths represent human navigation paths, collected by the game *Wikispeedia* in which a player navigates from one (given) article to another (given) article by following the links in the articles. Only paths reaching their determined target are considered. We model the Wikispeedia network as a directed, simple, unweighted graph (G_W) and the set of paths as $\mathcal{P}_W$.

Rush Hour This data set contains a state space of a single-player board game called *Rush Hour* where each node represents a possible game configuration and there is an edge from node v to node w if configuration w can be reached from v by a valid move in the game. We only include those nodes which are reachable from the node representing the start configuration of the game. Since all moves are reversible, the network is modelled as undirected unweighted graph (denoted by G_R) where one node represents the start configuration and one or more nodes represent configurations in which the game is solved (final nodes). A path in this data set is then the solution of a player trying to reach the final node from the start node by a sequence of valid moves. Only paths ending in a final node are considered. The set of paths will be denoted by $\mathcal{P}_R$. The data set was collected by Pelánek and Jarušek [12] by their web-based tool for education (available under tutor.fi.muni.cz).

DB1B This data set contains a sample of 10 % of all airline tickets of all reporting airlines, including all intermediate stops of a passenger's travel, provided by the US Bureau of Transportation Statistics which publishes for each quarter of a year the Airline Origin and Destination Survey (DB1B) [18]. We consider the passengers' travels for the years 2010 and 2011. A path is a journey of a passenger travelling from one airport to another, possibly with one or more intermediate stops. The network (G_D) is modelled as a simple directed and unweighted graph and is extracted from the ticket data by identifying city areas (possibly including more than one airport) as nodes, and adding a directed edge from a node v to a node w if at least one passenger's journey contains a flight from one airport in the area of node v to one airport in the area of node w. Passengers' journeys which are symmetric in the sense that the passenger travels from airport A to airport B over i intermediate airports and via the same intermediate airports back to A will be considered as two paths: one from A to B and one from B to A.

Data set	Source	Graph type	Nodes	Edges
Wikispeedia	[24,23]	directed	articles	hyperlinks
Rush Hour	[12]	undirected	configurations	valid game moves
DB1B	[18]	directed	cities	non-stop airline connections

Data set	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert\mathcal{P}\rvert$	path length	
				range	average
Wikispeedia	4592	119804	51306	$[1, 404]$	5
Rush Hour	364	1524	3044	$[3, 33]$	5
DB1B	419	12015	63681979	$[1, 14]$	1.3

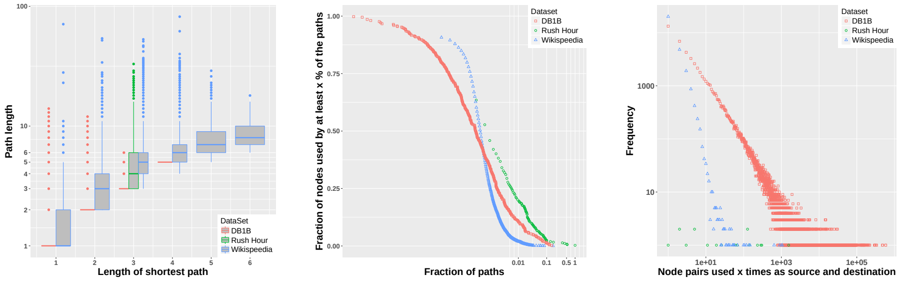
Table 1. Overview of the used data sets.

5 Assumptions of betweenness centrality

We chose our data sets such that the main assumption of the betweenness centrality are met: a process is flowing through the network by transfer mechanism, and whatever flows through it has a target to reach. Section 1 already pointed out that there are more assumptions. The next section will investigate whether those assumptions are satisfied in the selected data sets.

Process moves on shortest paths For each G_X and corresponding $\mathcal{P}_X$, $X \in \{W, D, R\}$, and for each $P \in \mathcal{P}_X$, we compare $|P|$ with $d(s(P), t(P))$. The results are shown in Fig. 2(a). For Rush Hour, $d(s(P), t(P)) = 3$ for each $P \in \mathcal{P}_R$ because we only consider paths reaching a final node from the start node—which is possible within three moves. For Wikispeedia, $1 \leq d(s(P), t(P)) \leq 6$ for all $P \in \mathcal{P}_W$. One outlier path of length 404 and two paths of length 101 and 104 have been taken out of the analysis. All node pairs in the DB1B network that were source and destination of any of the paths in $\mathcal{P}_D$ can be reached within 4 flights. We can observe that for the DB1B data set, the taken journey is almost always the shortest path between the airports. For Wikispeedia, the variation of the path lengths is much larger than for DB1B. Although the median length of the paths is greater than the length of the shortest path between its start and end node, the difference is only about 1. The same holds for Rush Hour. Hence, for all of the considered data sets, the assumption that the process is moving on shortest paths is approximately true.

Process moves between every node pair Since the betweenness centrality counts the (fraction of) shortest paths including a given node v and those fractions are summed up over all possible node pairs (s, t) (with $s \neq v$ and $s \neq t \neq v$), it is assumed that there is traffic between every pair of nodes. Table 2 shows whether this is a valid assumption for our data sets: In all data sets, almost all nodes are visited by at least one path. Since the network for DB1B was constructed from the paths, the number of used nodes is equal to the total number of nodes. It is not surprising that the fraction of used nodes in the Rush Hour data set is much smaller than in the other data sets, since all paths start at the network's



(a) Length of the paths vs length of shortest path from source to destination.

(b) Fraction of the nodes (y-axis) used by at least x % of the paths.

(c) Frequencies of node pairs used as source and target by at least one path.

Fig. 2. Analysis to which extent the given data sets satisfy the assumptions that the process moves on shortest paths (a), and that the process moves between each pair of nodes with the same intensity (c).

start node and end in one of the final nodes. In the Wikispeedia data set, 90 % of all nodes were used by at least one path. If, however, a node is only counted if at least ten paths include this node, the percentage drops to 74 %. Figure 2(b) shows this fact: it shows which portion of the nodes (contained in at least one path), is used by at least which portion of paths.

Although the majority of the nodes are included in at least one path of the data set, only a small fraction of the node *pairs* is actually used as source and destination of any path (cf. Tab. 2 and Fig. 2(c)). For Rush Hour and Wikispeedia, it rather seems to be an artifact: out of 132132 node pairs in G_R, only 20 pairs can actually be used by a valid solving path, since it must start in the start configuration and end in one of the 20 final configurations. In G_W, there are more than 21 million node pairs, but the data set contains “only” about $50k$ paths which explains the small value. For G_D on the other side, the fraction of used node pairs is, although the highest of all data sets, surprisingly small. Although the data contain $63m$ passengers’ travels over a time period of two years, almost half of all airport (city) pairs is never taken as a passenger’s journey.

Equal amount of flow between every pair of nodes While the previous paragraph showed that the assumption there is flow between every pair of nodes in the network, is not met in the data sets, this paragraph provides evidence that the assumption of equal flow between the nodes is also not true here. Figure 2(c) shows the frequency distribution of all node pairs which are the source and destination of at least one path in the data set, i.e., how many node pairs (y-axis) are used by exactly x paths in the data set as source and destination. Note the logarithmic scale on both axes. We see that although on different magnitudes, the behaviour is qualitatively the same for all data sets. When picking a node pair uniformly at random out of all node pairs used at least once, the probability is very high to pick a pair which is the source and destination of only one path, and very low to pick one which is the source and destination of more than thousand paths although there are more than $50k$ and $63m$ paths, respectively.

6 Process-driven betweenness centrality measures

Since we have shown that the usual assumptions of the betweenness centrality and about its underlying flow are not necessarily true, two questions arise. First, how much does it matter that those assumptions are not true, and second, can we do better? The information about how the entities move through the network is given in the data sets and can therefore be incorporated into a *process-driven betweenness centrality measure* (PDBC) which does not rely on the assumptions of shortest paths, flow between every node pair with equal intensity. The following section introduces four PDBC variants using different pieces of information contained in the data sets. The first two keep the assumption of using shortest paths (indicated by subscripted S), the next two count *real* paths from $\mathcal{P}$ (indicated by a subscripted R). All four require a graph $G = (V, E)$ and a set of paths $\mathcal{P}$ in G to be given. As a general framework, we introduce a weighted betweenness centrality by

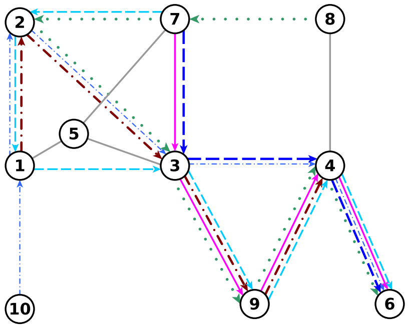
$$B_w(v) = \sum_{s \in V} \sum_{t \in V} w(s,t,v) \cdot \frac{\sigma_{st}(v)}{\sigma_{st}}$$

with a weight function $w : V \times V \times V \to \mathbb{R}$. The standard betweenness centrality is then B_c with the weight function $w_c(s,t,v) = 0$, if $s = t$ or $s = v$ or $v = t$, and $w_c(s,t,v) = 1$ otherwise, and the standard betweenness centrality including endpoints is B_E with the weight function $w_E(s,t,v) = 1$ for all $s,t,v \in V$.

Figure 3(a) shows an example graph with four paths in it, indicated by the directed edges of different colours. Then, $\mathcal{P} = \{P_1, P_2, P_3, P_4, P_5, P_6\}$ with $P_1 = (10,1,2,3,4,6)$, $P_2 = (7,3,4,6)$, $P_3 = (8,7,2,3,9,4,6)$, $P_4 = (7,3,9,4,6)$, $P_5 =$

Data set	Nodes	Used	Percentage
Wikspeedia	4592	4166	90%
Rush Hour	364	231	63 %
DB1B	419	419	100 %
Data set	Pairs	Used	Percentage
Wikspeedia	21081872	28706	0.14%
Rush Hour	132132	19	0.01 %
DB1B	175142	92242	52.7 %

Table 2. Usage frequencies of the nodes of the networks: first part of the table shows which fraction of the network nodes are used by at least one path in the data set. Second part shows which fraction of all possible node pairs are source and destination of at least one path in the data set.



(a) Example graph with paths.

Node	B_c	B_E	B_R	B_{SW}	B_S	B_{RW}
1	16.67	34.67	10	2	2	2
2	2.3	20.33	21.17	0	0	4
3	32	50	22	3.5	2.5	6
4	20	38	15	6	4	6
5	2.33	20.33	0	0	0	0
6	0	18	8	5	3	5
7	8.67	26.67	11	3	1	4
8	2	20	6	2.5	1.5	1
9	0	18	13.68	0	0	4
10	0	18	5	1	1	1

(b) Values of PDBC variants.

Fig. 3. Example graph G and PDBC values for the shown graph.

$(1, 2, 3, 9, 4)$, and $P_6 = (7, 2, 3, 9, 4, 6)$. The table in Fig. 3(b) shows the centrality values of the nodes with respect to each measure.

Note that the evaluation of the results in Sec. 7 will focus on the resulting *rankings* of the nodes with respect to the different measure variants, and not the actual measure *values*, hence, no effort is made with any normalization of the measures.

Variant B_S Keeping the assumption of shortest paths, this variant only considers shortest paths between nodes being source and destination of at least one path in $\mathcal{P}$. We define

$$B_S(v) = \sum_{s \in V} \sum_{t \in V} w_S(s, t, v) \cdot \frac{\sigma_{st}(v)}{\sigma_{st}}$$

with the weight function

$$w_S(s, t, v) = \begin{cases} 1 & \text{if } \exists P \in \mathcal{P} : s(P) = s \text{ and } t(P) = t \\ 0 & \text{else} \end{cases}$$

For Fig. 3, the weights are $w_S(1, 4, \cdot) = w_S(7, 6, \cdot) = w_S(8, 6, \cdot) = w_S(10, 6, \cdot) = 1$ and $w_S(s, t, v) = 0$ for all other $s, t, v \in V$. For node 3, the centrality value is then $B_S(3) = \frac{\sigma_{1,4}(3)}{\sigma_{1,4}} + \frac{\sigma_{8,6}(3)}{\sigma_{8,6}} + \frac{\sigma_{10,6}(3)}{\sigma_{10,6}} + \frac{\sigma_{7,6}(3)}{\sigma_{7,6}} = \frac{1}{1} + \frac{0}{1} + \frac{1}{1} + \frac{1}{2} = 2.5$.

Variant B_{SW} Section 5 showed that in the considered data sets, there is much more communication between some node pairs than between others. If a node is contained in most of the paths between all highly-demanded node pairs, the node should have a higher centrality than if it contained in the paths between less demanded node pairs. We therefore make the weight function proportional to the amount of flow between the node pair (hence the additional subsripted W). Formally, we define

$$B_{SW}(v) = \sum_{s \in V} \sum_{t \in V} w_{SW}(s, t, v) \cdot \frac{\sigma_{st}(v)}{\sigma_{st}}$$

with $w_{SW}(s,t,v) = |\{P \in \mathcal{P} | s(P) = s, t(P) = t\}|$. For the example in Fig. 3(a), this yields $w_{SW}(1,4,\cdot) = w_{SW}(8,6,\cdot) = w_{SW}(10,6,\cdot) = 1$, $w_{SW}(7,6,\cdot) = 3$ and $w_{SW} = 0$ in all other cases. Since shortest paths are counted, it is clear that often visited nodes as node 2 or 9 get a small value as they are not on any shortest path between the used nodes.

Variant B_R Unlike the previous two measures, this and the next variant count in how many *real* paths a node is contained in (therefore the subscript R). We define a *process-driven* version of σ_{st} and $\sigma_{st}(v)$: In order to keep the assumption that the process flows between any pair of nodes, we define $\sigma^{\mathcal{P}}_{\cdot st \cdot}$ as the number of paths in $\mathcal{P}$ *containing* s and t, and $\sigma^{\mathcal{P}}_{\cdot st \cdot}(v)$ as the number of paths in $\mathcal{P}$ that contain s and t, and v in between. (Otherwise, if the $\sigma^{\mathcal{P}}_{\cdot st \cdot}$ was defined as the number of paths in $\mathcal{P}$ with start node s and end node t, node pairs which are not start and end node of any path do not contribute to the centrality value.) We then define

$$B_R(v) = \sum_{s \in V} \sum_{t \in V} w_R(s,t,v) \cdot \frac{\sigma^{\mathcal{P}}_{\cdot st \cdot}(v)}{\sigma^{\mathcal{P}}_{\cdot st \cdot}}$$

with the convention $\frac{0}{0} = 0$ and $w_R(s,t,v) = 1$ for all $s,t,v \in V$ with $s \neq t$. Note that node pairs where at least one of the nodes is not contained in any paths in $\mathcal{P}$ do not contribute anything to the sum. In Fig. 3(a), all nodes except of node 5 are contained in a path from $\mathcal{P}$, we therefore get for node 2,

$$B_R(2) = \frac{\sigma^{\mathcal{P}}_{\cdot 1,2 \cdot}(2)}{\sigma^{\mathcal{P}}_{\cdot 1,2 \cdot}} + \cdots + \frac{\sigma^{\mathcal{P}}_{\cdot 7,9 \cdot}(2)}{\sigma^{\mathcal{P}}_{\cdot 7,9 \cdot}} + \cdots + \frac{\sigma^{\mathcal{P}}_{\cdot 10,6 \cdot}(2)}{\sigma^{\mathcal{P}}_{\cdot 10,6 \cdot}} = \frac{2}{2} + \cdots + \frac{2}{3} + \cdots + \frac{1}{1} = 21.17$$

We see that some nodes with a small centrality value with respect to B_c because they are not on (many) shortest paths, have a larger centrality value in this variant. For example, node 9 or node 2 have a minor importance with respect to B_S, but rise in importance wrt B_R because they are contained in a certain number of real paths.

Variant B_{RW} This variant combines all three kinds of information about the process in the network: counting real paths instead of shortest paths, only between those node pairs actually used as source and destination of the process instead of all, and weighting a node pair's contribution according to the amount of flow between the node pairs instead of assuming an equal amount. Formally, it is defined as

$$B_{RW}(v) = \sum_{s \in V} \sum_{t \in V} w_{RW}(s,t,v) \cdot \frac{\sigma^{\mathcal{P}}_{st}(v)}{\sigma^{\mathcal{P}}_{st}} = |\{P \in \mathcal{P} | v \in P\}|$$

with $w_{RW}(s,t,v) = w_{SW}(s,t,v) = |\{P \in \mathcal{P} | s(P) = s, t(P) = t\}|$. It is a kind of stress betweenness centrality. We see in Fig. 3 that node 3 which is central with respect to all previously considered centrality measures is also central with respect to B_{RW}, but nodes as 9 and 2 rise in importance because a considerable amount of paths passes through them.

Measures to compare We have introduced four variants of PDBC measures. Section 7 will describe how the centrality value of nodes (rather their position in the resulting ranking) will be affected when the PDBC measures are applied to the described data sets. For Wikispeedia and DB1B, it seems appropriate to compare the node rankings with those of B_E. However, for Rush Hour, this is different: since G_R represents the state space of a game, a valid path in this graph needs to start in the start node of G_R, and, in order to be a solving path, needs to end in one of the final states. This is why we introduce the *game betweenness centrality* for this case by

$$B_G(v) = \sum_{s \in V} \sum_{t \in V} w_G(s,t,v) \cdot \frac{\sigma_{st}(v)}{\sigma_{st}}$$

with $w_G(s,t,\cdot) = 1$ if s is start node and t final node, and $w_G(s,t,\cdot) = 0$ otherwise.

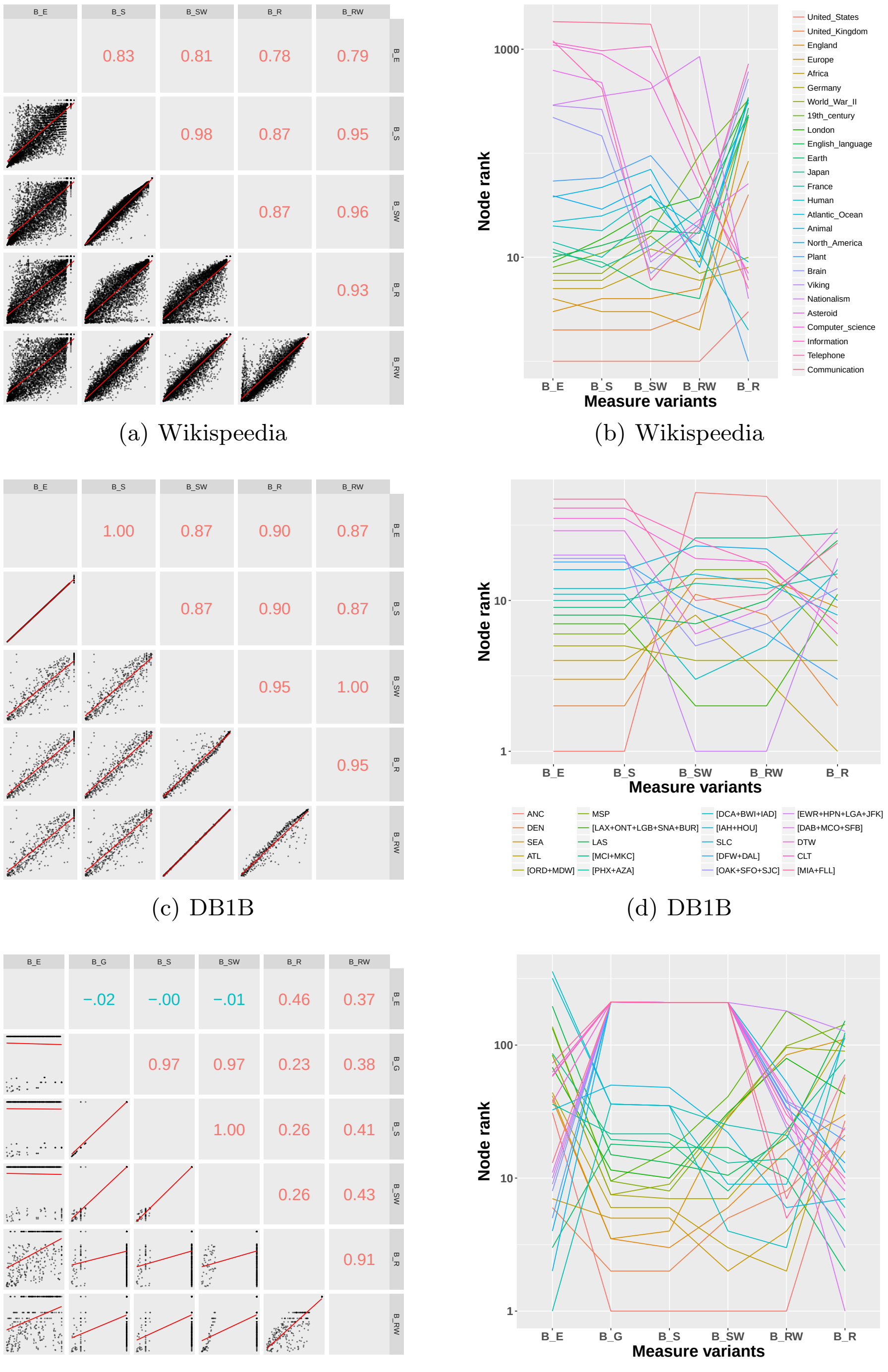
	COUNT	HOW	SUM OVER s, t WITH	WEIGHT
B_S	shortest paths		$\exists P \in \mathcal{P} : s \to t$	1
B_{SW}	shortest paths		$\exists P \in \mathcal{P} : s \to t$	$\#$ real paths $s \to t$
B_R	real paths	$\to s \to v \to t \to$	$s, t \in V$	1
B_{RW}	real paths	$s \to v \to t$	$\exists P \in \mathcal{P}$: $s \to t$	$\#$ real paths $s \to t$

Table 3. Categorization of the introduced process-driven betweenness centralities (PDBC).

7 Results

We computed the four PDBC variants and B_E for the networks described in Sec. 4. We are not interested in the exact values, but in the resulting order of importance of the nodes, only the resulting rankings are considered. We apply fractional ranking where nodes with the same measure value are assigned the average of ranks which they would have gotten without the ties.

Measure correlations Figure 4 (left) shows the rankings of the nodes in the data sets wrt all PDBC measures as well as B_E and (for Rush Hour) B_G. We can observe that for Wikispeedia and DB1B, there is a strong correlation between the node rankings wrt all variants. However, the correlations are less strong for Wikispeedia than for DB1B. Furthermore, the correlation of the PDBC variants with B_E is very low for the Rush Hour data set (largest value is 0.46). However, the rankings wrt the PDBC variants counting shortest paths (B_{SW}, B_S) correlate with the game betweenness centrality (correlation of 0.97), but not those variants counting real paths (B_R, B_{RW}). In all three data sets, although the most nodes have a similar ranking position in all variants, there are nodes which are rated as important by one measure, and very unimportant by to another measure.



(a) Wikispeedia

(b) Wikispeedia

(c) DB1B

(d) DB1B

(e) Rush Hour

(f) Rush Hour

Fig. 4. Left. Rankings of the network nodes in all PDBC variants and the standard betweenness centrality including endpoints (B_E). Low value on the axes indicates high centrality value. The red lines are linear regressions of the nodes' rankings. The value of the (rounded) Pearson correlation coefficient between the corresponding rankings can be found in the corresponding box, it is red if $p < 0.05$, blue otherwise. **Right.** Ranking behaviour of those nodes which are among the ten most central nodes with respect to at least one of the measures. Note the logarithmic scale on the y-axis. Colours are given according to the node's ranking wrt B_E (for Wikispeedia and DB1B) or wrt the game betweenness centrality (for Rush Hour).

Most central nodes We are interested in those nodes which are central with respect to at least one centrality measure. Figure 4 (right) shows the behaviour of the ranking of nodes which are among the ten nodes with highest centrality values for at least one of the measure variants. We can observe that in none of the data sets, it is the same set of ten nodes most central with respect to all measure variants. In all three data sets, the ranking wrt B_S is very similar to the "baseline" betweenness centrality, i.e., B_E for Wikispeedia and DB1B, and game betweenness centrality for Rush Hour. This implies that the PDBC variant with shortest paths and weights $\in \{0, 1\}$ does not affect the ranking of the most central nodes with respect to B_E. This is surprising because Sec. 5 showed that in those data sets, only a small part of all node pairs are actually used as source and target. Furthermore, for Rush Hour and DB1B, the ranking wrt B_{RW} is very different to the others, for Wikispeedia, however, the ranking with respect to B_{RW} is not too different to the ones wrt B_{SW}, and B_S. It is remarkable that for Wikispeedia, there is one node (the article *United States*) which is the most central with respect to B_E, but also the most (or at most third) central with respect to all PDBC variants. This cannot be observed in any of the two other data sets. Also the nodes of the Wikispeedia network on ranking position 2 to 5 wrt B_E are among the most central ones wrt all PDBC variants (except of B_R). The scale by which the considered nodes in- or decrease in ranking positions is different for the data sets: for DB1B, all nodes among the ten most central nodes wrt at least one of the variants, are among the 52 most central nodes wrt all variants (out of 419 possible ranking positions). This is different for the other two data sets: for Rush Hour, there are nodes which are among the ten most central nodes with respect to B_{SW} and B_{RW}, but have a ranking of 357 (out of 364) with respect to B_E. However, when examining those nodes, it turns out that those are final nodes and nodes adjacent to final states. More interesting are those nodes which are among the least central nodes with respect to the game betweenness centrality, but among the ten most central nodes wrt B_R and B_{RW}—those measures which count actually used paths instead of shortest paths. Those nodes are neither start nor solution states and although there are not on any shortest path from the start to any solution state, those nodes seem to be preferred by human players when solving this game. For Wikispeedia, there are nodes which ranking increases considerably: from ranking position 1843 (out of 4592) wrt B_E to position 5 wrt B_R (node *Communication*) or from position 1206 to 6 wrt B_{SW} (*Telephone*). Those big jumps in rank position might, however, partly be an effect of the data collection: there are four source-target-pairs which are suggested to the players with increased frequency as start and target node for the game for a certain period of time (*Pyramid*→*Bean*, *Brain*→*Telephone*, *Asteroid*→*Viking*, *Theatre*→*Zebra*) [22]. This means that paths with these sources/targets are contained more often than others. This explains why these nodes increase in importance when considering B_{SW}, but does not explain why they increase in importance in the same magnitude with respect to B_R.

8 Summary and Future Work

This work used available data sets containing the trajectories of entities in a network structure in order to compute a process-driven centrality measure for the nodes. The idea is to use the available information about how a process moves through the network in order to rank the nodes according to their importance with respect to this process. We chose data sets such that the two main assumptions of the betweenness centrality are met and it is therefore appropriate to actually use this centrality measure. We could show that while assuming shortest paths in those data sets can be justified, other assumptions of the betweenness centrality are not satisfied. We therefore introduced four variants of process-driven betweenness centrality measures (PDBC) which incorporate the information contained in the data sets. The resulting node rankings of most variants show high correlations to each other as well as to the standard betweenness centrality. Nevertheless, we can observe that incorporating the information about the process has an effect on the most central nodes with respect to the standard betweenness centrality and can in- or decrease a node's ranking by several thousands ranks. We furthermore observe that the different process-driven betweenness variants affect the node rankings in the three data sets in different ways.

Open questions left for future work are for example: it is obvious that the quality and validity of the results of PDBC is highly dependent on the quantity and quality of the available data. This yields two future directions: collect and compile more path data sets of high quality, and investigate which properties the set $\mathcal{P}$ must satisfy in order to make a reasonable statement. Additionally, networks develop dynamically over time: if an edge is not used by the process, it might disappear at some point, and if there is a large amount of traffic between two nodes with a larger distance, there might appear a shortcut between them. Considering the usage of the network by available path data might therefore a tool for predicting a network's change of structure.

References

1. Anthonisse, J.M.: The rush in a directed graph. Tech. Rep. BN 9/71, Stichting Mathematisch Centrum, Amsterdam (1971), (unpublished)
2. Bavelas, A.: A mathematical model for group structures. Human Organization 7(3), 16–30 (1948)
3. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2(1), 113–120 (1972)
4. Borgatti, S.P.: Centrality and network flow. Social Networks 27(1), 55–71 (2005)
5. Borgatti, S.P., Everett, M.G.: A graph-theoretic perspective on centrality. Social Networks 28(4), 466 – 484 (2006)
6. Carpenter, T., Karakostas, G., Shallcross, D.: Practical issues and algorithms for analyzing terrorist networks. In: Proceedings of the Western Simulation MultiConference (2002)
7. Dolev, S., Elovici, Y., Puzis, R.: Routing betweenness centrality. Journal of the ACM 57(4), 25:1–25:27 (2010)

8. Dorn, I., Lindenblatt, A., Zweig, K.A.: The trilemma of network analysis. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 9–14. Washington, DC, USA (2012)
9. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
10. Freeman, L.C.: Centrality in social networks conceptual clarification. Social networks 1(3), 215–239 (1978)
11. Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: A measure of betweenness based on network flow. Social networks 13(2), 141–154 (1991)
12. Jarušek, P., Pelánek, R.: Analysis of a simple model of problem solving times. In: Cerri, S., Clancey, W., Papadourakis, G., Panourgia, K. (eds.) Intelligent Tutoring Systems, Lecture Notes in Computer Science, vol. 7315, pp. 379–388. Springer, Berlin Heidelberg (2012)
13. Kleinberg, J.M.: Navigation in a small world. Nature 406, 845–845 (2000)
14. Koschützki, D., Lehmann, K.A., Peeters, L., Richter, S., Tenfelde-Podehl, D., Zlotowski, O.: Centrality indices. In: Brandes, U., Erlebach, T. (eds.) Network Analysis, Lecture Notes in Computer Science, vol. 3418, pp. 16–61. Springer Berlin Heidelberg (2005)
15. Milgram, S.: The small world problem. Psychology today 2(1), 60–67 (1967)
16. Newman, M.E.: A measure of betweenness centrality based on random walks. Social networks 27(1), 39–54 (2005)
17. Qin, J., Xu, J.J., Hu, D., Sageman, M., Chen, H.: Analyzing terrorist networks: A case study of the global salafi jihad network. Intelligence and security informatics pp. 287–304 (2005)
18. RITA TransStat: Origin and Destination Survey database (DB1B) (2016)
19. Rosvall, M., Esquivel, A.V., Lancichinetti, A., West, J.D., Lambiotte, R.: Memory in network flows and its effects on spreading dynamics and community detection. Nature communications 5 (2014)
20. Sabidussi, G.: The centrality index of a graph. Psychometrika 31(4), 581–603 (1966)
21. Sudarshan Iyengar, S., Veni Madhavan, C., Zweig, K.A., Natarajan, A.: Understanding human navigation using network analysis. Topics in cognitive science 4(1), 121–134 (2012)
22. West, R.: Human Navigation of Information Networks. Ph.D. thesis, Stanford University (2016)
23. West, R., Leskovec, J.: Human wayfinding in information networks. In: Proceedings of the 21st international conference on World Wide Web. pp. 619–628. ACM (2012)
24. West, R., Pineau, J., Precup, D.: Wikispeedia: An online game for inferring semantic distances between concepts. In: IJCAI International Joint Conference on Artificial Intelligence. pp. 1598–1603 (2009)
25. Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y.: T-drive: driving directions based on taxi trajectories. In: Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems. pp. 99–108. ACM (2010)
26. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World wide web. pp. 791–800. ACM (2009)