Appeared in "Finding Patterns of Human Behaviors in Network and Mobility Data (NEMO)", 2011

Augmented Betweenness Centrality for Mobility Prediction in Transportation Networks

Yaniv Altshuler[1], Rami Puzis[2], Yuval Elovici[2], Shlomo Bekhor[3], and Alex (Sandy) Pentland[1]

[1] MIT Media Lab
{yanival,sandy}@mit.edu

[2] Deutsche Telekom Lab, Department of Information Systems Eng., Ben-Gurion University
{puzis,elovici}@bgu.ac.il

[3] Transportation Research Institute, Technion
sbekhor@technion.ac.il

Abstract. Measuring and predicting the human mobility along the links of a transportation network has always been of a great importance to researchers in the field. Hitherto, producing such data relied heavily on expensive and time consuming surveying and on-field observational methods. In this work we propose an efficient estimation method for the assessment of the flow through links in transportation networks that is based on the Betweenness Centrality measure of the network's nodes. Furthermore, we show that the correlation between those two features can be significantly increased when additional (pre-defined and known) properties of the network are taken into account, generating an augmented *Mobility Oriented Betweenness Centrality* measure. We validate the results using a transportation dataset, constructed using cellular phones data, that contains a high resolution network of the Israeli transportation system. We show that the flow that was measured using this expensive and complicated method can be accurately estimated using our proposed Augmented Betweenness technique.

1 Introduction

The analysis of mobility trends and demands forecasting in transportation networks relies heavily on household survey data that provides the required input for calibrating the mathematical models that represent decisions people make related to travel [1]. However, a well known problem common to all interview-type surveys is non-response. Complex methods to correct for non-response have been developed, however, these alleviate the problem only partially [2].

As mentioned in [3], another limitation of household surveys is the need for active cooperation from the respondents, relying on their memory and patience. The need for active participation reduces the ability to capture complex travel and activity patterns, and the ability to collect data over a long period of time. The problems mentioned above, coupled with budget constraints, explain the fact that typical household surveys collect data regarding a period of merely one or two days for each household.

As a result, there exists a strong need for finding an alternative mechanism of assessing mobility and traffic demand in transportation networks, one that could be used without the necessary, tedious and inaccurate process of surveying.

Betweenness Centrality (BC) stands for the ability of an individual node to control the communication flow in the networks [4, 5]. Formally, for a node v it denoted the total portion of shortest-paths between every pair of nodes in the network that pass through v (see more details in Section 3). In recent years Betweenness was extensively applied for the analysis of various complex networks [6, 7] including among others social networks [8, 9], computer communication networks [10, 11], and protein interaction networks [12]. Holme [13] have shown that Betweenness is highly correlated with congestion in particle hopping systems. Extensions of the original definition of BC are applicable for directed and weighted networks [14, 15] as well as for multilayer networks where the underlying infrastructure and the origin-destination overlay are explicitly defined [16].

In this paper we discuss the applicability of BC and certain augmented types of it for the prediction of mobility patterns in transportation networks. Specifically, we show that there is a strong positive correlation between a traffic that flows through a node in a transportation network and its BC measures. In this study we use a comprehensive transportation network of the Israeli roads and highways system, containing over 15,000 directed links.

The rest of the paper is organised as follows :Section 2 describes the transportation data that was used in this study. Section 3 discusses the correlation between betweenness centrality and traffic flow, whereas concluding remarks appear in Section 4.

2 Transportation Network Dataset

The widespread use of cellular phones in Israel enables the collection of accurate transportation data. Given the small size of the country, all cellular companies provide national wide coverage. As shown in [3], the penetration of cellular phones to the Israeli market is very high, even to lower income households, and specially among individuals in the ages of 10 to 70 (the main focus of travel behavior studies). Such penetration enables a comprehensive study of travel behavior that is based on the mobility patterns of randomly selected mobile phones in the Israeli transportation system. This data was shown in [3] and [17] to provide a high quality coverage of the network, tracking 94% of the trips (defined as at least 2km in urban areas, and at least 10km in rural areas). The resulting data contained a wealth of traffic properties for a network of over 6,000 nodes, and 15,000 directed links. In addition, the network was accompanied with an Origin Destination (OD) matrix, specifying start and end points of trips.

The network was created for the National Israeli Transportation Planning Model. In urban areas the network contains arterial streets that connect the interurban roads. For each link of the network, there is information about the length (km), hierarchical type, free-flow travel time (min), capacity (vehicles per hour), toll (min), hourly flow (vehicles per hour), and congested travel time (min). The hourly flows and congested travel times were obtained from a traffic assignment model that loads the OD matrix on the network links.

2.1 Network Structure

Based on the dataset described above we have created a network structure, assigning running indices from 1 to 6716 to the nodes (junctions). We have examined the directed variant of the network where each road segment between two junctions was represented as either one or two directed links between the respective nodes.

In order to get a basic understanding of the network we first extracted and studied several of its structural properties (see Table 1). We have partitioned the network into structural equivalence classes of the nodes and bi-connected components and computed the betweenness centrality indices of the nodes [18, 19, 4]. Structurally equivalent vertices have exactly the same neighbors and the set of these vertices is called a structural equivalence class. As can be seen from Table 1 the number of structural equivalence classes is roughly the number of vertices in the network and the size of the largest class is three. This means that there are no "star-like" structures in the network and alternative paths between any two vertices are either longer than two hops or have other links emanating from the intermediate vertices. On the other hand the number of biconnected components in the network is low compared to the number of nodes, meaning that there are significant regions of the network that can be cut out by merely disconnecting a single node.

Table 1. Structural properties (Israeli transportation network).

Nodes	6716
Edges (undirected representation)	8374
Edges (directed representation)	15823
Number of structural equivalence classes	6655
Largest equivalence class	3
Number of bi-connected components (BCC)	931
Avg BCC size	8.2
Largest BCC	5778

2.2 Congestions

In this paper we define the impact of congestion as the difference between the time to travel through a congested link and the free-flow time to travel. Congestion of a junction can be either inbound or outbound. Inbound congestion is the sum of all congestions on inbound links of some junction. Figure 1 presents the distribution of congestion on network nodes (junctions). Power law nature of this distribution means that vast majority of nodes are not congested but there are a few nodes whose congestion can be arbitrarily large. Based on the *Wardrop's User Equilibrium* [20] this also implies a low number of yet significant deviations between the routes chosen by travelers during free-flow and during congestions. In Section 3.3 we use this fact to merge between two routing strategies.

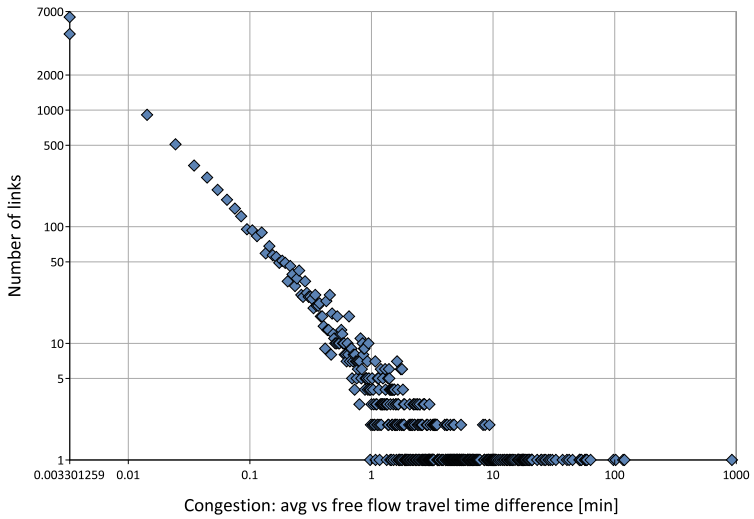

Fig. 1. Power law distribution of congestion.

2.3 Flow

The analyzed dataset contains traffic flow through links provided as the number of vehicles per hour. In the next section we will compare the flow through nodes estimated using Betweenness Centrality to the measured flow. We compute the total inbound flow through a node by summing flows on all of its inbound links, where outbound flow is computed symmetrically. Unless a specific junction is a source or a destination of traffic we expect the inbound flow to be equal to the outbound flow. Figure 2 demonstrates the correlation between inbound and outbound flow. We see that vast majority of the nodes are located on the main diagonal, however, there are some deviations, caused by the fact that the data represents average measurements that were carried out along a substantial period of time.

Figure 3 presents the distribution of inbound flow on network nodes. This distribution is exponential, meaning that a vast majority of nodes have little flow through them. However, in contrast to network congestion, there are no "unbounded fluctuations", i.e. the flow through the most "busy" junctions is not as high as can be expected from the power law distribution of betweenness and congestions (Figures 1 and 4). In fact, congestions significantly limit the flow through the busiest junctions, which subsequently is the reason we do not see the long tail in flow distribution.

3 Betweenness Centrality vs. Traffic Flow

Betweenness centrality is defined as the total fraction of shortest paths between each pair of vertices that pass through a given vertex [4]. Let $G = (V, E)$ be a directed transportation network where V is the set of junctions and E is the set of directed links

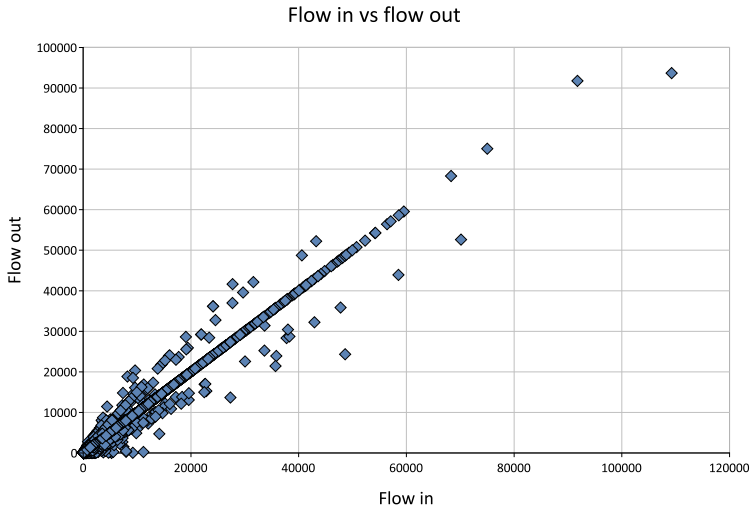

Fig. 2. Incoming vs. outgoing flow for each node.

as described in Section 2. Let $\sigma_{s,t}$ be the number of shortest paths between the origin vertex $s \in V$ and the destination vertex $t \in V$ (in some applications the shortest path constraint can be relieved to allow some deviations from the minimal distance between the two vertices). In the rest of this paper we will refer to the shortest or "almost" shortest paths between two vertices as *routes*. Let $\sigma_{s,t}(v)$ be the number of routes from s to t that pass through the vertex v. The Betweenness centrality can hence be expressed by the following equation:

$$BC(v) = \sum_{s,t \in V} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}. \tag{1}$$

Note that in this definition we include the end vertices (s and t) in the computation of Betweenness since we assume that vehicles can be inspected also at their origin and at the point of their destination.

After computing the Betweenness centrality for the given transportation network, we can easily see that the distribution of Betweenness centrality follows a power law (Figure 4). Long tail distributions such as the power law suggest that there is a non negligible probability for existence of vertices whose Betweenness centrality can be arbitrarily high. This is in contrast to the exponential flow distribution depicted in Figure 3. The different nature of these two distributions suggests that BC as defined above will overestimate the actual traffic flow through nodes especially for the most central vertices.

Next we would like to check the correlation between BC and traffic flow. Although the correlation is significant the square error is very low ($R^2 = 0.2021$) as shown in Figure 5 (a). Every point in this Figure represents a vertex with the x-axis corresponding to the measured traffic flow and y-axis corresponding to the computed BC.

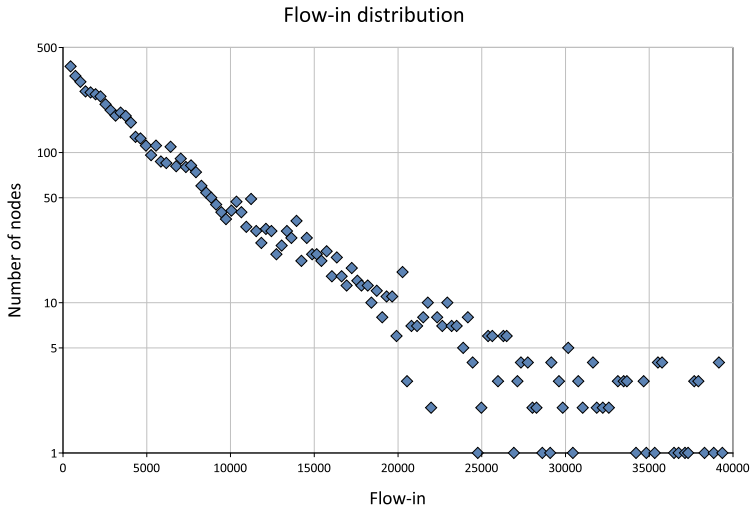

Fig. 3. Exponential distribution of traffic flow through nodes.

We now discuss augmented variants of the Betweenness centrality measure that significantly improve the correlation with the traffic flow.

3.1 *Origin-Destination* based Betweenness Centrality

BC definition according to Equation 1 BC assumes equal weights of routes between every pair of vertices in the network. In other words every vertex acts as an origin and as a destination of traffic. We would like to utilise the measured origin-destination (OD) flow matrix in order to prioritize network regions by their actual use. For this, we shall use the following altered definition for betweenness, as suggested in [16]:

$$BC(v) = \sum_{s,t \in V} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}} \cdot OD_{s,t} \tag{2}$$

where OD is the actual measured origin-destination matrix. This method produces a better correlation ($R^2 = 0.4916$) between the theoretic (BC) and the measured traffic flow (see Figure 5 (b)).

3.2 Shortest Routes based on Time to Travel

In order to further improve our ability to estimate the predicted network flow using the network's topology, we note that both BC calculation methods (Equations 1 and 2 above) assume that routes are chosen according to shortest path strategy based on hop counting. In this section, we retain the shortest path assumption but use weighted links for calculating the Betweenness score. One option is to use the length of the road

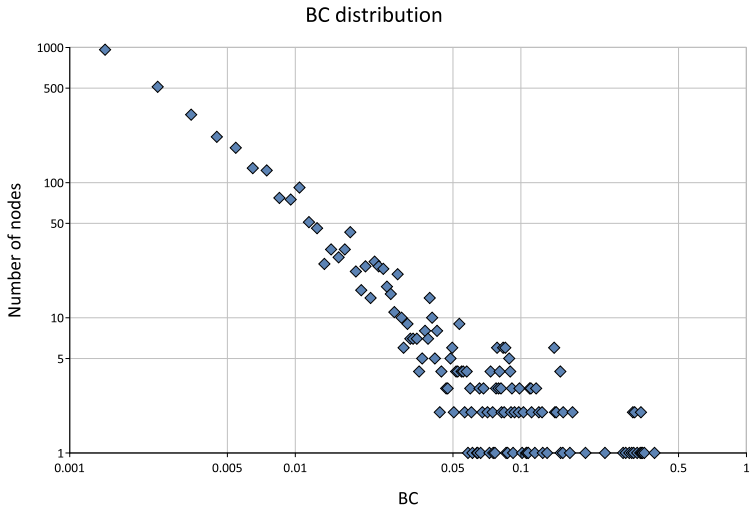

Fig. 4. Power law distribution of Betweenness centrality

segments as their weights for the shortest path calculations (based on the well justified assumption that people prefer short routes over the long ones). However, the road capacity, congestions, and the number of segments also play significant roles when choosing the route to destination. People would prefer highways over sideways when the distance difference is not high.

Shortest path algorithms (such as Dijkstra's or Bellman-Ford's) are able to consider only one distance weight on links when computing the shortest path to a destination. We shall therefore assume that the primary heuristic guiding people when they chose a route is the time required to reach their destination. Using this assumption, we recompute the BC on the directed transportation, weighting links by their free-flow travel time.

Let $BC^{ft}(v)$ denote the Betweenness of a node v computed w.r.t. the free-flow travel time. Figure 5 (c) shows significant improvements in the correlation between the measured traffic flow and the theoretical BC^{ft} values computed w.r.t the OD matrix and free-flow travel time link weights ($R^2 = 0.6123$). We can see that there are few nodes whose flow was significantly underestimated by the BC measure. Notice that there are also several nodes whose flow was actually overestimated. This can be explained by the fact that people do not travel strictly via shortest paths, but may have various deviations. In particular the deviations form shortest paths are affected by the day time and the day of week.

3.3 *Peak-Hours Aware* Betweenness Centrality

It is a reasonable assumption that during peak hours travelers will choose to avoid the congested roads and choose their routes based on the congested travel time rather than on the free-flow travel times. Let $BC^{ct}(v)$ denote the Betweenness of a node v com-

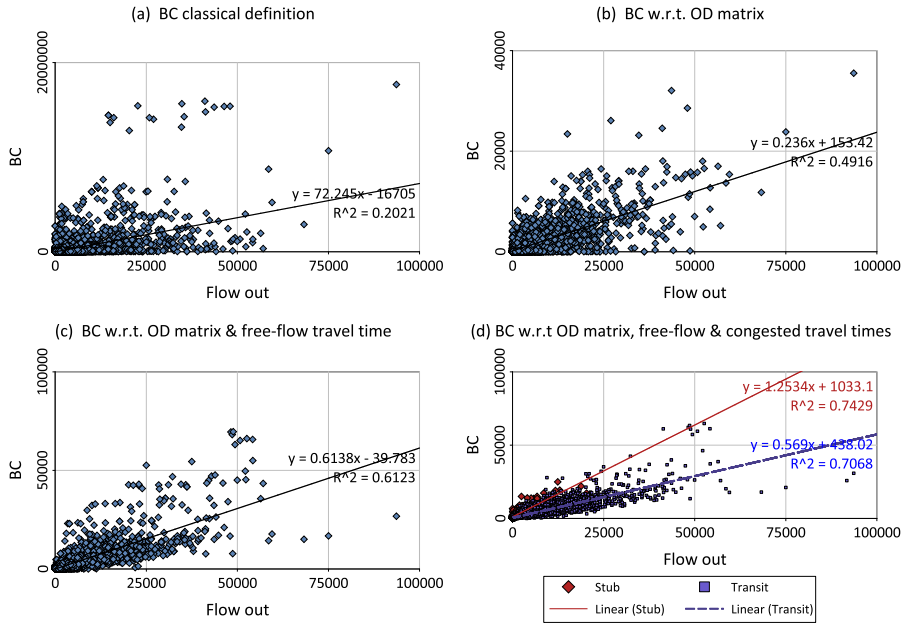

Fig. 5. Correlation of flow through nodes and Betweeness Centrality

puted w.r.t. the congested time. Computing Betweenness using only the congested travel time weights results in $R^2 = 0.7096$. Although peak hours are relatively small fraction of the day, most vehicles travel at these hours. This is the reason for higher correlation of BC^{ct} with the measured traffic flow.

We shall now combine both the Betweenness centrality computed w.r.t. the free-flow travel time and the congested time by taking a weighted average, namely :

$$BC(v) = \alpha \cdot BC^{ft}(v) + (1 - \alpha) \cdot BC^{ct}(v)$$

where α denotes the relative fraction of vehicles traveling during the free-flow periods. The resulting centrality index can achieve higher correlation with the measured average traffic flow. The maximal correlation of $R^2 = 0.7285$ is obtained for $\alpha = 0.25$ as shown in the Figure 6.

3.4 Separating *Stubs Nodes* from *Transit Nodes*

Carefully looking at the various nodes we can see that they can be divided into two groups : *stub nodes* and *transit nodes*.

A Stub node is a node that is an origin or a destination of the traffic (as seen in the Origin-Destination matrix). These nodes account for approximately 10% of the network's nodes. All other nodes (namely, nodes that generate insignificant or no outgoing or incoming routes) are called Transit nodes, as they only forward traffic and do not generate or consume it.

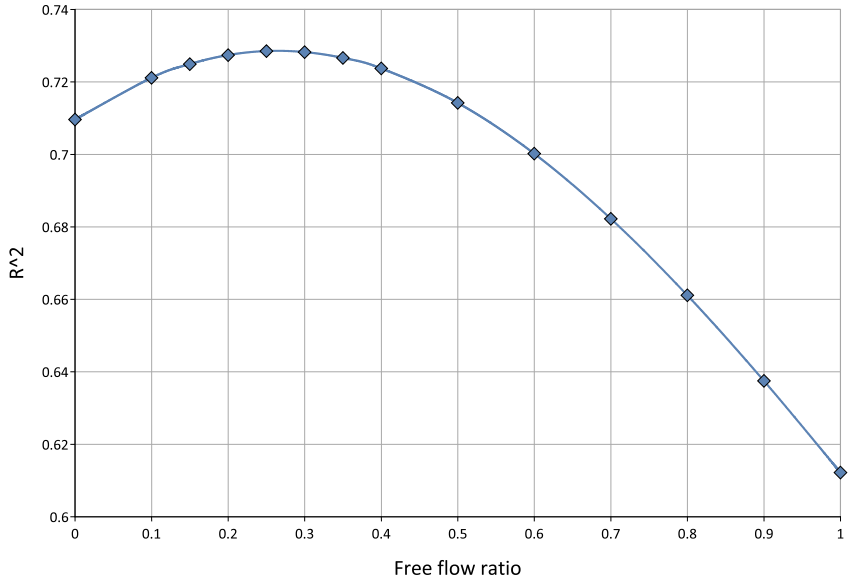

Fig. 6. Squared error (R^2) as the function of the free flow traffic fraction (α).

Figure 5 (d) presents the correlation that is received when the two groups of nodes are being processed separately. Specifically, the results show a $R^2 = 0.7068$ for the Transit nodes and a $R^2 = 0.7429$ for the Stub nodes.

3.5 *Mobility Oriented* Betweenness Centrality

As previously mentioned, the transportation network dataset we use contains a "*type*" attribute for each link, representing the domain-specific "role" of the link in the overall network. For example, links of types 13 and 14 correspond to internal neighborhood roads, whereas links of type 12 correspond to "collectors" — roads that are in charge of aggregating the traffic from neighborhood roads and channeling it to metropolitan roads, and so on. As each type of roads have therefore a different role, we now try to further improve our flow prediction by examining the Betweenness values achieved when calculating it for every group separately.

The results of the correlation that is achieved using this method are presented in Figure 7. We can clearly see that for the more important roads (namely, those with lower type number, representing a more infrastructurial role in the transportation network) this technique yields R^2 values that are consistently above 0.74, reaching 0.83(!) for road of types 2 and 9 (note that roads of type 90 are fictive roads with infinite capacity that were artificially added in order to connect distinct regions in the network).

It should be noted that each node may have incoming roads of different types. Each plot corresponds to a set of nodes whose max incoming road type is as specified. In addition, the BC calculations were not made for each set of nodes separately — BC was

computed for the complete network, while the correlations were computed separately for each type.

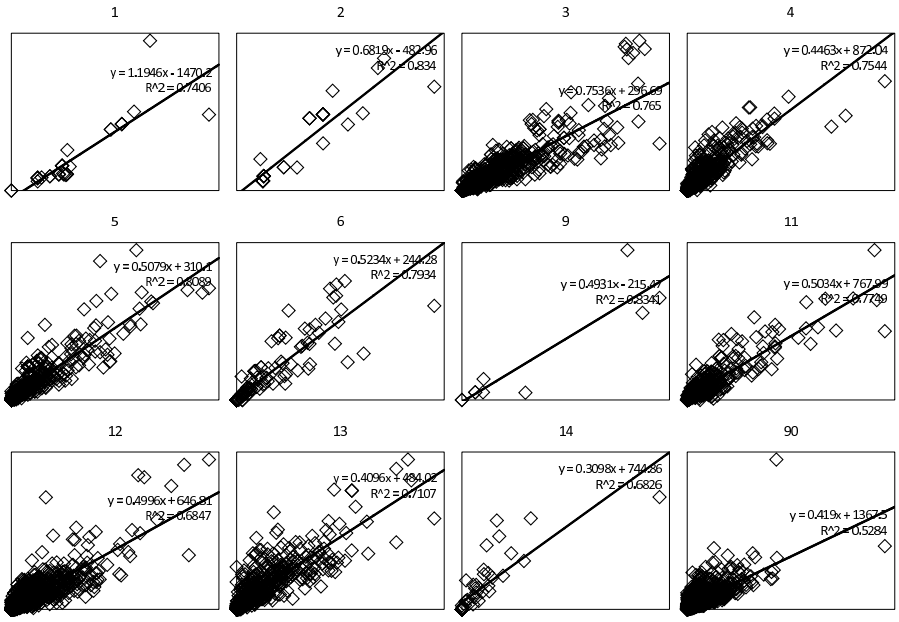

Fig. 7. Correlation of flow through nodes and Betweenness (computed separately for different *types* of links.

4 Conclusions

In this paper we have discussed the correlation between the Betweenness centrality of a node and its expected traffic flow, in transportation networks. Using a comprehensive dataset that covers the Israeli transportation network we have first performed a simple analysis of the network and its properties, showing that there exists a correlation between the traffic flow of nodes and their Betweenness centrality. We then revised the basic definition of Betweenness centrality, showing that when analyzing the network in a way which takes into account additional known properties of the links (specifically, time to travel through links), a much stronger correlation can be achieved. Taking into account that a large portion of the traffic is being generated during rush hours, and that different roads may have different 'roles' in the transportation network, we show that a significantly higher correlation can be achieved when clustering the roads into groups based on their types (a known property of each road), while also giving increased weight to data that is associated with certain hours. Using this method that we call "*Mobility*

Oriented Betweenness Centrality" we demonstrate correlation values of approximately $Z^2 = 0.8$.

This method can now be used in order to generate highly accurate approximations of the traffic flow in the network, based on its topology, the OD matrix, and time to travel without costly simulations. Furthermore, we can also use this method in order to estimate the dynamic changes in traffic flow due to changes in the Betweenness of nodes, caused by events such as car accidents, road detours, etc. This technique can be useful for traffic prediction systems, such as *DynaMIT* [21].

In addition, based on the correlation between individual flow and Betweenness flow, a similar correlation between *Group Betweenness* and group flow can be implied. Subsequently, various problems dealing with *flow* that are relatively hard to solve can now be tackled using their dual Betweenness problems. For example, a knap-sack style problem of finding the best group of nodes to put speed cameras at (in order to capture as many speeding drivers as possible) can be translated to a dual problem of finding a group of nodes with the largest group Betweenness. For the latter, however, there exist various efficient approximation heuristics, that can be used in order to derive a solution for the first. Similar approach was taken in [22, 23] for optimizing deployment of traffic inspection systems in communication networks.

References

1. P. Stopher, C.G. Wilmot, C.C. Stecher, and R. Alsnih. Household travel surveys: Proposed standards and guidelines. *Travel Survey Methods. Quality and Future Directions*, 2006.
2. A.J. Richardson. Data structures, sampling and survey issues: Report of workshop m6. 9th international association of travel behaviour research conference, (australia). 2001.
3. S. Bekhor, Y. Cohen, and C. Solomon. Evaluating long-distance travel patterns in israel by tracking cellular phone positions. *Journal of Advanced Transportation*, pages n/a–n/a, 2011.
4. L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.
5. J. M. Anthonisse. The rush in a directed graph. Technical Report BN 9/71, Stichting Mathematisch Centrum, Amsterdam, 1971.
6. S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, March 2001.
7. M. Barthélemy. Betweenness centrality in large complex networks. *The European Physical Journal B – Condensed Matter*, 38(2):163–168, March 2004.
8. S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge, England: Cambridge University Press., 1994.
9. J. Scott. *Social Network Analysis: A Handbook*. Sage Publications, London, 2000.
10. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Comm. Rev.*, 29(4):251–262, 1999.
11. S.H. Yook, H. Jeong, and A.-L. Barabasi. Modeling the internet's large-scale topology. *Proceedings of the National Academy of Science*, 99(21):13382–13386, Oct. 2002.
12. P. Bork, L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee, and E. M. Marcotte. Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.*, 14(3):292–299, 2004.
13. P. Holme. Congestion and centrality in traffic flow on complex networks. *Advances in Complex Systems*, 6(2):163–176, 2003.
14. D. R. White and S. P. Borgatti. Betweenness centrality measures for directed graphs. *Social Networks*, 16:335–346, 1994.

15. U. Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.
16. R. Puzis, M. D. Klippel, Y. Elovici, and S. Dolev. Optimization of nids placement for protection of intercommunicating critical infrastructures. In *EuroISI*, 2007.
17. Y.J. Gur, S. Bekhor, C. Solomon, and L. Kheifits. Intercity person trip tables for nationwide transportation planning in israel obtained from massive cell phone data. *Transportation Research Record: Journal of the Transportation Research Board*, 2121:145–151, 2009.
18. F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1):49–80, 1971.
19. J. Lerner. *Network analysis: methodological foundations*, chapter Role Assignments. Springer LNCS 3418, 2005.
20. WARDROP J. G. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, 1:325–378, 1952.
21. M. Ben-Akiva, M. Bierlaire, H. N. Koutsopoulos, and R. Mishalani. *Real-time simulation of traffic demand-supply interactions within DynaMIT*. Transportation and network analysis: current trends. Miscellenea in honor of Michael Florian. Kluwer Academic, 2002.
22. R. Puzis, Y. Elovici, and S. Dolev. Finding the most prominent group in complex networks. *AI Comm.*, 20:287–296, 2007.
23. S. Dolev, Y. Elovici, R. Puzis, and P. Zilberman. Incremental deployment of network monitors based on group betweenness centrality. *Inf. Proc. Letters*, 109:1172–1176, 2009.