WHAT DOES SOCIAL MEDIA SAY ABOUT THE INFRASTRUCTURE CONSTRUCTION PROJECT?

Mazdak Nik Bakht, PhD Candidate, mazdak.nikbakht@mail.utoronto.ca
Tamer E. El-diraby, Associate Professor, tamer@ecf.utoronto.ca
Department of Civil Engineering, University of Toronto, Ontario, Canada

ABSTRACT

The process of public consultation for planning and construction of the urban infrastructure as a sociotechnical system requires a bidirectional interaction and dialogue among all stakeholders of the project. Policy makers, official and technical decision makers, and the public should get together in form of a social network to discuss different aspects of the project. Social media and the social Web can play the role of a platform to accommodate a social network and keep the flow of related discussions. Detecting cores of interest formed in such networks and profiling stakeholders of the project (including the end users) can help decision makers in the process of demand detection, public engagement, and marketing the project to the public community. This requires topology analysis of the social connections and semantic analysis of the discussions. This paper introduces some initial steps in this regard. It combines community detection algorithms with information retrieval practices into a hybrid technique to detect and profile the communities of the project followers and cores of interest in the network of stakeholders of the urban infrastructure project. This technique is used to analyze micro-blogging website Twitter for cross-town LRT project to profile the followers based on their interests and the ideas they support. Analysis of the content and sentiment of the discussions is currently an underway research and would be discussed elsewhere.

Keywords: ifrastructure discussion network, scial network analysis, oline social media, cmmunity detection, iformation retrieval

1. INTRODUCTION

Civil Infrastructure in the modern society is a complex network of physical assets together with the actors who develop, maintain, and operate them, or use the services provided by the developed system. This network cannot be functional without both of the physical/technical components, and the mishmash of the social interactions and interdependencies among them. As endorsed in the literature (Parkin (1994), Kroes, et al. (2006), and Ottens, et al. (2006) among others), such characteristics redefine the civil infrastructure as a sociotechnical system (rather than a pure technical artifact, or an engineering system). Decision making for construction of such a system must happen within a network rather than a hierarchy of decision makers (Bruijn & Heuvelhof , 2000), and should involve as many social interactions among the stakeholders (including the end users) as possible. As the result, '*public relation*' programs in infrastructure projects are evolving into an '*engagement partnership*'; i.e. instead of updating the public through a top-down transfer of information from the project management team to the community, the community engagement programs today target other forms of consultation by involving the public through a two-way communication. On the other hand, social connectivity and epidemiology of the knowledge over the Web 2.0 (social web) is incorporating producers and consumers of the 'knowledge' into the "*prosumers*" (a portmanteau formed by contracting the word professional, or also producer, with the word consumer); today, it is becoming more and more difficult to draw a bold line between the two. In the context of decision making for infrastructure, the citizens are becoming important sources of input for development and maintenance plans, and can be better defined as prosumers rather than mere users. The prosumerism trend has

recently been the source of new business models and scientific achievements. Websites such as Amazon, Google, and Facebook are making money out of it, and a project such as Wikipedia collect corpus of the knowledge, relying on the power of prosumers. Governments and other macro level decision makers in the AEC industry have also started to benefit from the prosumers culture in the process of engagement partnership for infrastructure projects. As the result, only in North America, 82 out of the 100 strategic infrastructure projects (announced by North America strategic infrastructure leadership in 2011) have active Facebook or Twitter accounts to complete the bidirectional communication path between official/technical decision makers and nontechnical/community decision contributors. With more than 550 million followers, and still growing at the average rate of 135,000 new users per day, micro-blogging website Twitter records 58 million tweets every day. People express their opinion about many different issues – including the built environment – in less than 140 character statements, and this can be a significant opportunity for decision makers to detect the citizens' demands/feedback and to communicate with them.

In this sense, Web 2.0 plays the role of a platform which not only brings the decision makers and decision contributors together and connects them to each other in form of a heterogeneous network, but also documents and showcases their ideas and keeps the flow of project-related discussions among them. As the result, a combination of people and ideas about the construction project are linked to each other as explained by El-Diraby (2011), and form an *Infrastructure Discussion Network –IDN* (Nik Bakht & El-diraby, 2013). Members of the knowledge-enabled e-society use the free access to the information about the construction project and discuss its different aspects, related decisions, and alternative solutions amongst themselves. Analysis of the social connectivity and the content of discussions on the IDN can provide the decision makers with a significant opportunity in the process of network decision making for the infrastructure project. Detecting influential actors (to be targeted for consultation in marketing the project decisions and reverse marketing for the feedback and demands), as discussed by Nik Bakht and El-diraby (2013) is one of these opportunities. As determined by the theory of network decision making, the final decision should always be in form of a 'package deal' to address the majority of interests in the project (Bruijn & Heuvelhof , 2000). Detecting cores of interest in followers of the project is another important insight that analysis of IDN can provide the decision makers with. This can be done either by a semantic clustering of the ideas discussed, or through community detection among the followers. While this paper is discussing the latter method, studies on the former approach are currently underway and will be addressed somewhere else.

This paper is focusing on detecting the communities of followers in an IDN, and profiling the core interests in each community based on the commonalities and particularities of that community. In this respect, first of all, different algorithms for community detection in the SNA are briefly introduced and compared against some performance measures to evaluate their applicability to the context of IDN. After selecting efficient community detection algorithms, the question would be: how to classify the main interests in each community of the IDN? This paper tailors methods benchmarked from document management and information retrieval to answer this question. The tools will then be applied to the actual case of an LRT development project in Toronto-Canada to detect the core interests of the project followers and the terms discussed among them. Once the sentiment of the discussions is added, this can provide the decision makers with the profile of the social opinion about the project and the social interests that must be addressed by the final decisions.

2. BACKROUND LITERATURE AND RELATED WORKS

A 'network' can be generally defined as a set of interdependent individuals (nodes) with all of the interactions among them (edges or links). If the nodes are social entities (actors), and the edges are social linkages (relational ties or social interactions), then the resulted network is called a 'Social Network'. Social network analysis (SNA) is a very well established line of research in many domains and has found extensive applications during the past decades. Specifically, with the prevalence of the social web, SNA has found many applications in the analysis of online social networks. On the other hand, the construction industry, has been organized in project networks (which as Taylor and Levitt (2007) suggest, are social networks of experts working together on specific construction projects) since late 1950's (Stinchcombe, 1959). Pryke (2004) suggested the network perspective of the construction project as a more mature structure than the traditional hierarchical management, and the social

network model of construction was introduced by Chinowsky, et al. (2008) to increase the efficiency in the management of projects through the management of nontechnical components and the team performance. This model had two main components: dynamics (to address dynamics of interactions), and mechanics (to address the free flow of knowledge among project participants). Nik Bakht and El-diraby (2013) suggest that for urban infrastructure, this network should b e extended to include the end users in form of the IDN (as discussed).

Social networks are typically composed of groups of nodes which are tightly connected among themselves and are sparsely connected to nodes from other groups. This behavior has roots in the formation process of the network and is due to the fact that people are more likely to join the communities in which not only they have more friends, but also the friends are more densely connected to each other. The densely connected cores are called '*clusters*', '*modules*', or '*communities*'. Community detection is in fact the problem of graph partitioning to find such densely connected clusters. Various classes of community detection algorithms have been developed in the SNA literature, some of which are briefly introduced here.

Strength of weak ties and community detection – This algorithm is proposed by Girvan and Newman (2002) and can be used for un-weighted, undirected graphs. It divides the graph into hierarchical clusters using the notion of edge betweenness (number of shortest paths in the graph which pass through an edge) and based on the fact that the local bridges of a graph have high centrality. The algorithm simply works by finding weak ties (edges with the highest betweenness centrality) and removing them. As the result, any new connected component will be taken as a new community. The process is recursively repeated until no edge is left.

Modularity maximization – Modularity maximization is the most popular method for finding cores of the social networks. Modularity is defined as the difference between number of existing edges in a partition and expected number of edges which can exist among the nodes of that partition. Modularity of a partition (Q) can be estimated as (Newman , 2006):

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (1)$$

In which :

- A_{ij}: is the ij entry of the graph's adjacency matrix (weight of link ij)
- k_i: is the degree of node i
- m: is the total number of elements
- c_i: is the community that node i belongs to, and
- $\delta(u, v)$: is the Dirac delta function

The higher the modularity of a partition is, the denser that community will be. Therefore, problem of community detection can be formulated as partitioning a graph into groups such that the summation of modularity over all partitions is maximum. It is shown that this would be an NP-hard optimization problem with no direct/explicit solution. However, various heuristic methods or theoretical approximation algorithms are suggested in the literature for solving such a combinatorial maximization problem. *Agglomerative clustering* (Clauset, et al., 2004), *Fast modularity optimization* (Newman , 2006), and *Multi-level aggregation* (Blondel, et al., 2008) are examples of such methods.

Spectral Graph partitioning – This is an optimization problem of dividing vertices of the graph into groups in the format that maximizes number of connections within the groups and minimizes the number of connections between the groups. Shi and Malik (1997) achieve this goal by defining the ratio of bridges between partitions to the total edges in the two partitions as the objective function of a minimization problem. Leskovec, et al. (2008) introduce 'conductance' (or normalized cut metric) as the number of edges pointing outside the community divided by the number of edges inside the community, and detect the optimum partitioning by minimizing the conductance over all communities.

Trawling – Trawling is the name of a method proposed by Kumar, Raghavan, and colleagues (1999) for detection of cyber-communities over the Web, based on the shared interests (e.g. subscription to the similar webpages). This method is founded on a theorem in random graph theory which states that every large enough/dense enough bipartite random graph $G(X,Y,E)$ (With edge set E, and node sets X and Y), with a high probability contains a core $K_{s,t}$ (a complete bipartite graph among s and t number of nodes which are completely connected to each other and completely disconnected among themselves). Based on this argument, the problem of

detecting communities in bipartite graphs can be reinterpreted as the problem of finding complete connected cores which can be known as the 'signature' of communities.

Clique percolation – In most of the real world/large scale social networks, communities overlap and there are nodes that lie at the intersection of multiple communities. CPM (Clique Percolation Method) is used to detect such communities. CPM works based on detection of maximal '*cliques*' and '*k-clique communities*' in a graph. A *k*-clique is a complete (fully connected) subgraph of size *k* and a set of adjacent *k*-cliques form a *k*-clique community. The CPM algorithm involves in two main steps: finding maximal cliques (cliques in the graph which cannot be expanded to cliques of higher *k* values) (Tomita, et al., 2006), and merging overlapping maximal cliques found to detect the *k*-clique communities (Palla, et al. 2005).

Community detection has found extensive applications in various domains including biology, social science, bibliometrics and scientific collaboration, marketing, etc. Even providing a list of all these applications would be beyond the scope of this paper; however, applications such as topic detection in collaborative tagging systems (as reviewed by Papadopoulos, et al. (2012)), and social trust evaluation for recommender systems (addressed by Pitsilis, et al. (2011)among others) pinpoint that community detection can lead into profiling and grouping users based on their common interests. This is due to the simple fact: 'birds of the same feather flock together', which indirectly is reflected in the topology of the social connectivity graph. Communities of a network can be thought of as groups of people getting together around a common interest, and therefore in the context of the IDN, detecting the communities can help to distinguish some of the interests around the construction project. Some potential applications of community detection in the IDN for the official decision makers of the infrastructure project can be listed as:

- Profiling the users (or future users) of the facility/system which is being developed,
- Profiling the cores of interest around the project,
- Finding community leaders to be engaged by the public engagement programs,
- Detecting the interactions and interrelations among different communities,
- Finding the possible bottlenecks in the process of public communication and partnership for infrastructure projects.

Achieving these goals would not only depend on in detecting communities in form of clusters of the IDN graph, but also requires 'labeling' them based on the common interests which exist among the members of each community. It also requires monitoring the social discussions within and between the communities. Direct search for the shared interests and clustering the nodes based on the similarity of their interests is one solution to handle this issue. Steinhaeuser and Chawla (2008) suggested to define the interests as nodes' attribute and then to cluster the graph based on the attribute similarity among nodes. In a similar study, Kalafatis (2009) harvested the Twitter users based on their similar interests by looking for occurrence of some pre-defined keywords in their biography on the twitter. Although applying such methods can profile and classify the main interests around the project, they ignore the 'networkedness' of the interests and ideas discussed. Various interests/ideas around a project not only get weight due to the number of people who support them, but also (and maybe more importantly) based on how densely these supporters are connected to one another. We refer to such a property as the *network value* of the interests/ideas expressed over the IDN. Evaluation of the interests within the context of their network values requires an inverse procedure; i.e. detecting the topological communities, and then mining the nodes' interests within each community to find the similarities among the community members. This paper suggests the latter approach by combining community detection algorithms and document retrieval methods. The methodology is explained in the next part.

3. COMMUNITY DETECTION AND LABELING/PROFILING COMMUNITIES IN THE IDN

Various metrics and measures can be considered to assess the efficiency of different community detection algorithms and to select a suitable algorithm for the IDNs. In general, algorithmic efficiency, scalability, stability, and accuracy can be named as the most popular measures of performance for combinatorial and computational graph algorithms. *Efficiency* is mainly related to the computational complexity of the algorithm and is usually expressed in terms of the resource requirements: execution time, and memory usage. *Scalability* is a measure of increase in amount of resource usage for the algorithm as the dimensions of a problem grow. *Stability* of the

algorithm on the other hand, is related to the stability of the results each time the algorithm is applied to the same problem, under the similar conditions. In the scope of community detection, various metrics are introduced in the literature to measure the *accuracy* of results which include coverage, conductance, modularity, and Normalized Mutual Information (NMI). These are numerical measures for quality of the community structure that the algorithm produces (for detailed definition of each, one can see Moradi, et al. (2012) and Lancichinetti & Fortunato (2010)). Although scalability is a challenge to most of the graph mining algorithms; as most of the algorithms discussed in this paper are originally developed to handle large-scale social networks (including tens of thousands to millions of nodes), the scalability of the algorithms would not be a critical feature for the scope of IDNs (which are typically medium-scale networks composed of dozens to thousands of nodes). However, as some of the community detection algorithms use heuristic (and sometimes meta-heuristic) optimization techniques, each time the algorithm is applied, the results might slightly change, and therefore, the stability of the algorithm should be investigated.

Finding the most suitable algorithm(s) for the particular context of IDNs (as medium-scale, technical issue-centered networks) cannot be limited to analysis of the computation performance; rather, a more detailed study on real-world cases, followed by a qualitative analysis and evaluation of the result – as suggested by Papadopoulos, et al.(2012) – would be necessary. Such studies are currently underway and the results will be reported elsewhere. However, most of the studies to date prove efficiency and competency of algorithms working based on modularity maximization. Table 1 gives an overview of the results of studies on performance of the community detection methods. For the purpose of the current study, we take it from here and test different variants of such algorithms (the results will be discussed in the next part). In the following, these algorithms are explained in more details.

Table 1: Community detection algorithms reviewed by this paper

Algorithm Class	Performance Measure			Recommended for
	Computational Performance	Stability	Scalability	
Strength of weak ties	Low	Stable	Low	Un-directed – Un-weighted Graphs
Modularity Maximization	High	Depends on optimization algorithm	High	All graphs
Spectral Graph Partitioning	High	Stable	Depends on algorithm	All graphs
Trawling	Average	Stable	Average	Detecting the 'signature of networks' in large-scale–dense bipartite graphs
Clique percolation	Low	Depends on the features	Average	Detecting overlapping communities

Newman (2006) Fast Modularity Optimization Method: This algorithm partitions the graph into two communities at each step, and continues iteratively until no more splitting is allowed. Whenever the graph [or a community of it] is divided into two groups, a Boolean vector $\boldsymbol{s}$ (called '*community membership vector*') is defined for each group in which any entry corresponds to one node of the graph [or the community], and the value of the entry is +1 if the node belongs to the first group and is –1 if the node belongs to the second group. Therefore, each iteration of the algorithm can be formulated as looking for vector $\boldsymbol{s}$ which maximizes the overall modularity after splitting, and is followed through three major steps: *Step 1– Modularity Matrix*: Calculating a matrix **B** for the graph which is called modularity matrix, entries of which are calculated as:

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \tag{2}$$

Step 2– Leading eigenvector: Finding the leading (first) eigenvector of Matrix **B** (called $\boldsymbol{u}_1$) using numerical techniques such as power method; *Step 3– Splitting by rounding rule*: Splitting the nodes of the graph based on the sign of elements of $\boldsymbol{u_1}$, i.e.:

$$s_i = \begin{cases} +1 \; ; \; if\ i'th\ entry\ of\ \boldsymbol{u}_1\ is\ positive \\ -1 \; ; \; if\ i'th\ entry\ of\ \boldsymbol{u}_1\ is\ negative \end{cases}$$

The procedure stops when all communities are indivisible (all the entries of u_1 's in each community are either positive or negative), or the overall modularity does not increase anymore by splitting.

Clauset, Newman , & Moore (CNM 2004) Agglomerative clustering: This is a bottom-up methods which instead of starting from the large graph and hierarchically partitioning it, starts by forming small communities via local optimization of modularity, and then aggregates communities to form bigger clusters. Algorithm can be explained in three steps: *Step 1* – Putting each node in its own community; *Step 2* – Computing (change in the modularity if communities *x* and *y* are joined) for each pair of communities; *Step 3* – going over all communities and merging the pairs which result in the highest . The algorithm stops when no further reduction in is possible.

Blondel, et al. (2008) Louvain method: This is another bottom up algorithm with a similar nature to CNM, working based on heuristic optimization and folding. Initially, each node in the graph belongs to its own community (therefore a graph of size N would have N communities at the beginning). Then an iterative greedy algorithm is applied such that at each step all nodes are visited following a standard order, and the neighbors are merged as long as the modularity is increasing. At the end of each iteration all nodes in the same community are merged into one 'hyper node', and weighs are assigned to the edges between these hyper nodes to represent the number of intercommunity connections. This procedure which is called 'folding' ensures rapid decrease in the number of nodes that must be examined at each step, and increases the algorithm efficiency.

Once the communities of the IDN are detected, a modified version of Term Frequency–Inverse Document Frequency (TF-IDF) is applied to the discussions and/or biographies of the nodes in each community, to detect the dominating themes of opinions and interests for each community. TF-IDF is usually used in text mining for topic detection. TF-IDF is in fact a term weighting system composed of two components: TF, which gives higher weight to the terms with higher occurrence in a text (as they are more likely to be descriptive than the terms with low frequency in the document), and IDF, which scores down the common terms in multiple documents (as they are less likely to be good discriminators for a particular document). Therefore, if term *i* appears f_{ij} times in document *j*, then:

$$TF_{ij} = {}^{f_{ij}}/_{m_j} \qquad (3)$$

in which $m_j \max_i(f_{ij})$, and if *n* shows the number of documents, then:

$$IDF_i = \log\left({}^{n}/_{1 + d_i}\right) \qquad (4)$$

where d_i is the number of documents in which term *i* has occurred (document frequency). d_i is added to one to prevent zero at the denominator. Several experiments have demonstrated that the product of the two components (called TF-IDF) for a particular word *i*, with respect to a particular document *j* gives a measure of how descriptive the word can be for the document. By taking each community of the IDN as one document including all the discussions stated by the nodes in that community, a TF-IDF analysis can result in detecting terms which uniquely describe each community. Details of the technique are explained through an example in the next part.

4. CASE STUDY

The Eglinton-Scarborough "Crosstown" (a light-rail system) is one of the largest transit projects currently underway in North America. It is a part of the bigger city-wide transit plan called "Transit City", which was announced in 2007 and has been under long debates since then. Cancellation of Transit City by the mayor (who was supporting a subway alternative) in the late 2010 and resuming it again in the early 2012 by Toronto City Council and under several causes including the pressure of the community (mainly in form of a social movement called: 'Save Transit City'), are among other reasons which put this project under the spotlight of the social attention. Crosstown is an $8.2 Billion, 25.2 kilometer east-west LRT line passing through a congested corridor of the midtown of Toronto and is running underground in major parts (19.5Km). The street-level segment is planned to be separate from the street traffic with raised medians. TTC (Toronto Transit Commission) is the eventual owner and operator of the project. Metrolinx (a Provincial planning and finance agency) is the owner's agent (and the general contractor) in construction. Several Canadian contractors and consultants are the other "technical"

partners of the project. Procurement began on March 2011with manufacturing of the pre-cast tunnel linings, and the opening is planned for 2020. The construction officially launched on November 2011, and the tunneling is expected to begin by summer 2013.

The history of urban transit plans in the city proves the high cost of social opposition, as well as the high sensitivity of the community in the city of Toronto regarding transit projects. Therefore, several community meetings for public consultation have been held and more are planned on different aspects of the project (such as general specifications, station designs, construction schedule, and operation plan). In the era of online social media, official decision makers also launched a Twitter profile (with screen name: 'CrosstownTO') for the project on December 2011. At the time data was collected (September 2012), CrosstownTO had 521 followers, and in less than year, this number has increased into more than 1000 followers. The network of CrosstownTO followers is the example of an IDN. In the following, we apply the method proposed in this paper to detect the communities of followers, and to profile them based on two parameters: profile description, and the recent tweets posted by the followers. We take the short biography in the user's profile (in this paper called '*bio*') as a description of affiliation and concerns of the person, and the theme of users' tweets as an indicator of the users' interests.

4.1. Data collection and analysis methodology

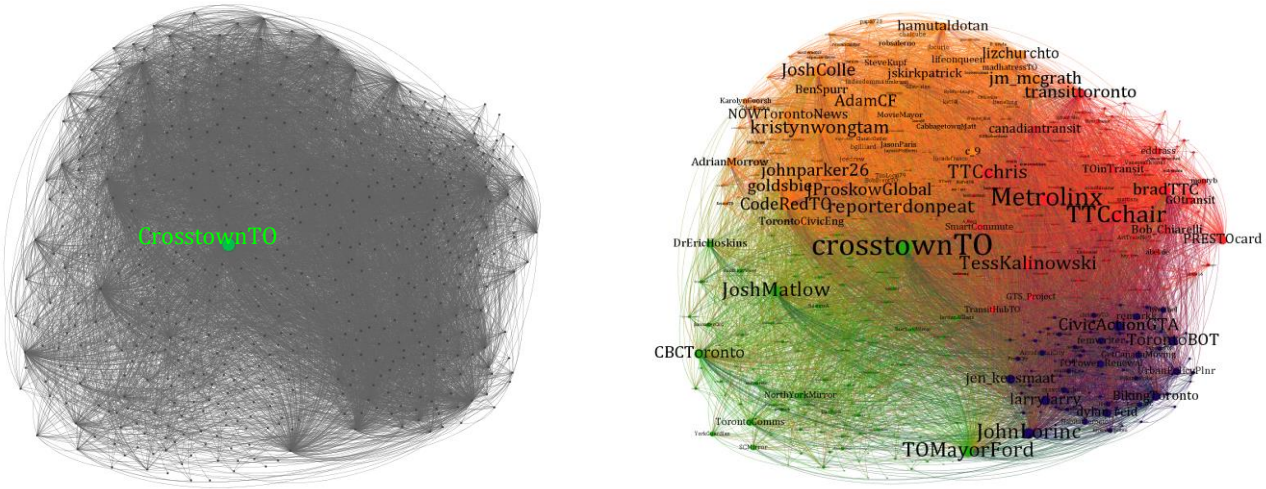
Connectivity among followers of the Crosstown project has been detected by collecting data from Twitter API (Application Programming Interface). The result, as of August 2012, is the graph shown in Figure 1(a), where each node represents a Twitter profile and a directed edge from node A to node B indicates that A is following B on Twitter. Version 0.8.1 of Software *Gephi* (https://gephi.org/) is used to visualize the network. This is an 'ego-centered' network having 'CrosstownTO' as its focal actor. Ego-centered networks consist of a focal actor (termed ego) and a set of alters who have ties to it. Such networks are widely used by anthropologists to study the social environment surrounding individuals or families, as well as by sociologists to study the social support. Collecting data for users with protected profiles is not possible. However, as the main focus of this study is 'influence', it can be simply assumed that while tweets by such individuals are not publically accessible, these actors cannot have a high impact on the general network. Therefore, in the current network, such profiles are nodes of the graph with only one outgoing link (following the project profile only).

Different modularity maximization algorithms discussed in the previous part, as well as the Girvan-Newman hierarchical clustering algorithm are applied to the network and the results are compared against the performance measures. In order to analyze community shared interests and the content of discussions among the communities, the TF-IDF analysis is applied to the users' descriptions (bio's), and their last 50 tweets. The texts are normalized into ASCII standard and transformed into lower-case (to minimize spelling differences), and the non-descriptive terms (such as the numbers, the gibberish, punctuations, etc.) have been removed using a standard stop list, augmented by some additional terms. URLs, and mentions (@username) are also removed and two dictionaries are generated: one for the terms used in the users' bio (which contains 1,913 terms), and one for the last 50 tweets (having 33,135 words). The TF-IDF is applied at the community level, i.e. each community is considered as a document composed of its users' descriptions (or the tweets by members of the community). Term frequencies in each community, and the inverse document frequency of the terms among the four communities are calculated and combined to result in the TF-IDF of all the terms used by the community members.

The standard TF-IDF method is originally developed for cases with numerous documents ($n \rightarrow 1$), and needs to be slightly modified to become applicable to the case of limited communities ($n = 4$). The +1 in the denominator of equation (4) suggests that if $d_i = n/1$ (i.e. the term is repeated in all except one of the communities), then the IDF for such a term would be zero, and the term will be removed (irrespective of its frequency). In our case we need to correct the formula to get a small (yet non-zero) value for when $d_i = n \mid 1$. One solution can be using . As all terms in this problem are at least happening in one of the communities, and we will not have division by zero problem. However, we still need to be able to track the terms which are occurring in all the four communities. In order to achieve this goal, we have modified the IDF as follows:

$$IDF_i = \log\left(n/0.05 + d_i\right) \quad (5)$$

This helps to always get positive/nonzero values for the IDF, except when a term is common among all the communities in which IDF would return a negative, still nonzero number. Therefore, if the TF-ITF for a term in a community is zero, it shows that the term has had TF=0, or in other words it has never been used by the users of that community.



(a) IDN is an ego–centered graph

(b) Communities of the IDN detected by fast unfolding (color of a node shows the community it belongs to)

Figure 1: Network of 'CrosstownTO' twitter ID followers as an example of an IDN

4.2. Analysis results

Results of applying community detection algorithms are summarized in Table 2. Given the size of IDNs in general, the execution time would not be the most critical metric to compare the performance of algorithms; however as it is seen, the Girvan-Newman algorithm (working based on the betweenness centrality) has considerably lower efficiency and accuracy. Computational performance of the modularity maximization algorithms do not have a substantial difference. Although the fast unfolding method gives slightly different outputs every times it is applied, the highest accuracy is resulted from this method: the graph is clustered into four communities, shown in different colors by Figure 1(b), and the total modularity is 0.181.

Table 2: Results of applying community detection algorithms to the CrosstownTO IDN

Algorithm	# of communities detected	Performance Measure		
		Efficiency (Execution Time)	Accuracy (Modularity)	Stability
Edge betweenness (Girvan-Newman)	2	>30 min	0.001	Stable
Agglomerative clustering (Newman-Moore)	5	<5 Sec	0.1701	Stable
Louvain fast unfolding (Blondel et al.)	4-6	<5 Sec	0.169 –0.181	Unstable

A closer look at each community reveals interesting details about the social structure behind the followers of Crosstown LRT project. Nikbakht and El-diraby (2013) have analyzed the top influential nodes of each community and show that the first community, shown in green by Figure 1(b), is mainly under influence of the politicians (such as the mayor and some provincial ministers). Nodes of the second (red) and the third (orange) communities are respectively dominated by the technical/official decision makers of the project (TTC-the owner, and Metrolinx-the general contractor), and some policy makers of the city government (mainly city councilors). The fourth community (in dark blue) is the community of the public, in which the influential nodes are some journalists (city hall and transit reporters in particular), urban planners, and transportation experts with no official

affiliation to the project. This has been done in a semi manual manner by manually going through the description of the top influential nodes after detecting them automatically within each community. The TF-IDF analysis can be an automatic method to confirm those results. As illustrated in Table 3, particular themes can be detected for users' descriptions in each community. High frequency of hashtags with different combinations of TTC in the second community, terms such as strategist, ward, and councillor in the third community, and terms such as author, journalist, city hall, magazine, and Torontoist (a city blog about Toronto) clearly show the dominant themes which bring the people together in each of these communities. However, given the broadness of the concepts covered by the politicians and their dependents' biographies, there is a relative diversity in theme of terms with high TF-IDF in the first community.

Table 3:Top descriptive terms in bio of the users at each community

	Community			
	C1	**C2**	**C3**	**C4**
Terms with high TF-IDF	university	Go	waste	journalist
	Eglinton	public	comms	cityhall
	Davisville	area	energy	freelance
	industry	construction	strategist	Torontoist
	town	Hamilton	sustainable	author
	updates	dedicated	ward	civic
	air	#TTC	economics	magazine
	adventure	GTA	RyersonU	write
	business	#TTChelps	ceo	neighborhood
	TV	#TTCnotices	councillor	drink
	Politicians	**Technical/Official Decision makers**	**Policy Makers**	**Public**

Taking such an analysis one step further and mining the topics discussed by users in each community can give decision makers a better idea about the followers' opinions, concerns, and interests. Table 4 presents the relevant words of high TF-IDF from the last 50 tweets by the users of each community. This can be known as a summary index of the social discussions about this project, clustered based on the community of people who support them. Particularly, this table can help to track the communities who have or have not used specific terms; a positive number shows that the term has been used, zero shows that the term has never occurred in the tweets by the community members, and a negative number means that all the four communities have used the term in their tweets. In general, this table shows that no term exists in the project-related context which is exclusively used by one community and not the others. the hashtag *#St.clair-disaster-glorified-streetcars-subway* refers to another LRT project in the city of Toronto, which eventually due to the lack of clear communication with the public ended up to a lawsuit against the city by the local businesses and the neighborhood community (TTC, 2010). All groups of followers of CrosstownTO, except the project technical and official decision makers have referred to this experiment. The public (community C4) is interested in almost all different topics discussed, this can be due to the presence of reporters and journalists in this community. *Monorail* (as an alternative to the LRT system), used by the politicians, decision makers, and city policy makers has never been repeated by the public. The public has instead used the term *monorails* (in plural form) a few times, which seems to be related to benchmarking other cities with monorail system. It is also interesting that all four communities except the politicians have used the hashtag *#fix-transit-now*. Hashtag *#develop-transit-etc*. on the other hand has not been used by the politicians and policy makers. #transitcity and #antitransit have been the words common between the public and the politicians, which can refer to the long negotiations against and in favor of the transit city project.

Focusing on the bottom part of the table shows that terms such as: *subway*, *highways*, and *transit-system*, which mainly reflect the debates on other alternatives and substitutions for the crosstown LRT, as well as *preconstruction*, *contracting*, and *carbon-tax* which are generally related to the current phase of the project (at the time the data was collected) have been used by the users from all the four communities. This gives a general sense about the inter-communities social discussions about the project. Figure 2 is an illustration of the terms shared by different communities in their most recent 50 tweets.

Table 4: Social dialogues; within and between communities opinion exchanges through analysis of the users' last 50 tweets and detection of relevant terms with high TF-IDF

Term	C1	C2	C3	C4
monorail	1.49E-02	5.90E-03	9.22E-04	0.00E+00
Lrts	2.98E-03	1.47E-03	9.22E-04	0.00E+00
contract	0.00E+00	1.47E-03	9.22E-04	2.95E-03
light-rail	0.00E+00	1.47E-03	1.84E-03	9.82E-04
fix-transit-now	0.00E+00	1.47E-03	9.22E-04	9.82E-04
monorails	0.00E+00	2.95E-03	1.84E-03	9.82E-04
transit	0.00E+00	7.37E-03	9.22E-04	9.82E-04
Toronto-transit	0.00E+00	4.42E-03	9.22E-04	2.95E-03
St.clair-disaster-glorified-streetcars-subway	2.68E-02	0.00E+00	9.22E-04	9.82E-04
Unsustainable	2.98E-03	0.00E+00	9.22E-03	3.93E-03
construction10	2.98E-03	0.00E+00	9.22E-04	9.82E-04
freight	7.34E-03	0.00E+00	2.27E-03	0.00E+00
Transitcity	7.34E-03	0.00E+00	0.00E+00	1.21E-02
anti-transit	7.34E-03	0.00E+00	0.00E+00	4.84E-03
commuters	0.00E+00	3.63E-03	6.82E-03	0.00E+00
development-transit-etc.	0.00E+00	3.63E-03	0.00E+00	7.27E-03
clean-air-commute	0.00E+00	3.63E-03	0.00E+00	2.42E-03
Lrt	0.00E+00	3.63E-03	0.00E+00	2.42E-03
traffic	0.00E+00	3.63E-03	0.00E+00	2.42E-03
trains	0.00E+00	3.63E-03	0.00E+00	2.42E-03
reconstruction	0.00E+00	7.27E-03	0.00E+00	9.69E-03
renewal	0.00E+00	7.27E-03	0.00E+00	9.69E-03
streetcar	0.00E+00	1.09E-02	0.00E+00	7.27E-03
revitalisation-reconstruction	0.00E+00	0.00E+00	4.54E-03	4.84E-03
preconstruction	-1.36E-04	-2.03E-04	-4.22E-05	-9.00E-05
carbon-tax	-2.73E-04	-4.05E-04	-3.80E-04	-5.40E-04
highways	-2.73E-04	-4.05E-04	-3.80E-04	-9.90E-04
contracting	-4.09E-04	-7.43E-04	-1.27E-04	-2.70E-04
subway	-8.18E-04	-3.38E-04	-1.69E-04	-2.25E-04
transit-system	-1.23E-03	-1.49E-03	-9.29E-04	-1.49E-03

5. CONCLUSION AND FUTURE WORK

Development of infrastructure as a sociotechnical system requires a wider range of public decision contributors to be involved in the process of decision making in form of an Infrastructure Discussion Network (IDN). Such networks will have a bottom-up and chaotic nature. This paper offers a combination of mathematical methods and information retrieval techniques for detecting and labeling/profiling the communities in the mishmash of the IDN. We introduced the community detection algorithms available in SNA and compared their applicability against some performance measures. We concluded that given the general size and nature of the IDNs, modularity maximization has the satisfactory performance in detecting communities of IDNs. As the communities are usually formed around common interests, we benchmarked methods from text mining to automatically label the communities' shared interests. It was shown that such analyses can also detect the keywords in the social network discussions and reveal the patterns of communication existing among different groups of stakeholders.

What was presented here can create the guidelines showing who must be involved in the engagement and negotiation processes, together with their concerns, and the interests which must be addressed by the final decision. In combination with a proper sentiment analysis of the online discussions, findings of this paper can provide the official decision makers with a mental map of the major project followers. This method has the eminence of being self-organizing and emerging out of interactions of the actors from within the system. Real interactions in the offline world are more or less reflected through the online behavior of the actors. This can provide a powerful tool for communication process with the public. Also, since offline social opposition most of

the time lags the online declaration of dissatisfaction, detecting online alarms can give the official decision makers enough time to change the decisions appropriately, or to apply timely policies to prevent formation of snowballing social opposition and to reduce the risk of failure in such projects before it is too late.

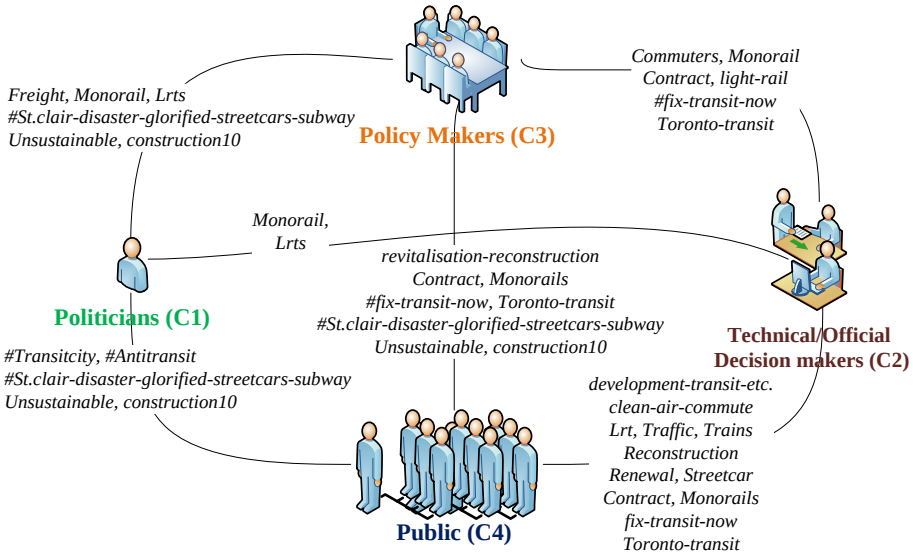


Figure 2: Co-occurrence of the terms in the last 50 tweets of the CrosstownTO IDN communities

Finally, the limitations of the method should be admitted. Twitter (or any other social media) does not necessarily represent an exact picture of the society. There will always exist people who have a high impact on their communities but are not active in online environment or do not have a Twitter account. Furthermore, the online, and offline social attitudes do not necessarily match perfectly all the time. Consequently, existence of a followership relation between two nodes on the Twitter might not essentially be a notion of the real social relational tie. The ongoing research is focused on closer types of connection and stronger ties among the followers of projects (such as mentioning and re-tweeting). On the other hand, communities of the IDN can be known as an incomplete/imperfect image of the communities of interest in the real project. However, as IDN's mature this imperfection tends to the perfection. Moreover, TF-IDF stops at the term level and does not go beyond the topic. Although providing decision makers with the list of key terms can give an overview of the social concerns and interests; detecting the content and sentiment of the discussions is necessary for finding the clear layout of the social mental map about the project. In the studies which are currently underway, natural language processing is tried to handle this issue. Also the issue of confidentiality of the project-related as well as followers-related information can be another barrier resulting in the online discussion outline being a distorted picture of the reality.

While the idea of combined network of official and non-official decision makers is a futuristic view which needs more experiments over the time to settle and stabilize, what was presented in this paper is applicable into any other platform in which connectivity of people discussing construction of infrastructure is recorded, and it has the notable advantage of detecting patterns of order in the chaos of IDN, and interpreting them.

REFERENCES

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Kournal of statistical mechanics: Theory and experiment, 10*, 2000-2012.

Bruijn, H. d., & Heuvelhof , E. t. (2000). *Networks and decision making.* Utrecht: LEMMA Publishers.

Chinowsky, P., Diekmann, J., & Galotti, V. (2008). Social network model of construction. *ASCE Journal of construction engineering & management*.

Clauset, A., Newman , M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 066111-1 to 066111-6.

El-Diraby, T. E. (2011). Civil infrastructure as a chaotic socio-technical system: How can information systems support collaborative innovation. *CIBW078.* Computer Knowledge Building.

Girvan, M., & Newman, M. E. (2002, June). Community structure in social and biological networks. *Proceecings of national academy of sciences of the USA, 99*(12), 7821-7826.

Kalafatis, T. (2009, May). *Twitter analytics: cluster analysis reveals similar twitter users.* Retrieved from Life analytics: http://lifeanalytics.blogspot.ca/2009/05/twitter-analytics-cluster-analysis.html

Kroes, P., Franseen, M., Van de Poel, I., & Ottens, M. (2006). treating socio-technical systems as engineering systems: sme conceptual problems. *System research adn behavioral science, 23*, 803-514.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the web for emerging cyber-commuinities. *Computer networks: The international journal of computer and telecommunications networking*, 1491-493.

Lancichinetti, A., & Fortunato, S. (2010). Community detection algorithms: a comparitive study. *Physical review E 80*.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. (2008). Statistical properties of community structure in large social and information networks. *WWW08 Social netwokrs and web 2.0* (pp. 695-704). Beijing, China: IW3C2.

Moradi, F., Olovsson, T., & Tsigas, P. (2012). An evaluation of community detection algorithms on large-scale email traffic. *Lecture notes in computer science, 7276*, 283-294.

Newman , M. E. (2006). Modularity and community structure in networks. *PNAS (Proceedings of the National Academy of Sciences of United States of America*, 8577-8582.

Nik Bakht, M., & El-diraby, T. E. (2013). Analyzing infrastructure discussion networks: order of 'influence' in chaos of 'followers'. *Csce annual conference-4th construction specialty conference* . Montreal: CSCE .

Ottens, M. M., Franssen, M., Kroes, P. A., & Poel, V. I. (2006). Modelling infrastructures as socio-technical systems. *International journal of critical infrastructures*, 133-145.

Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 814-818.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *J. data mining and knowledge discovery, 24*(3), 515-554.

Parkin, J. (1994). A power model of urban infrastructure decision making. *Ceoforum*, 203-211.

Pitsilis, G., Zhang, X., & Wang, W. (2011). Clusteing recommenders in collaborative filtering using explicit trust information. *Proceedings of: Trust Management V* (pp. 82-97). Copenhagen, Denmark: Springer.

Pryke, S. D. (2004). Analysing construction project coalitions: exploring the application of social network analysis. *Construction management and economics, 22*(8), 787-797.

Shi, J., & Malik, J. (1997). Normalized cuts and image segmentation. *IEEE Computer Vision and pattern recognition* (pp. 731-737). IEEE.

Steinhaeuser, K., & Chawla, N. V. (2008). Community detection in a large real-world social network. *International conference on social computing, behavioral modeling and prediction* (pp. 168-175). Phoenix, Arizona, USA: Springer.

Stinchcombe, A. (1959). Bureaucratic and craft administration of production: A comparative study. *dministrative Science Quarterly, 4*(2), 168-187.

Taylor , J., & Levitt , R. (2007). Innovation alignment and project network dynamics: an integrative model for change. *Project management journal*.

Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical computer science*, 28-42.

TTC. (2010). *Transit city implementation- the St.Clair project experience.* Toronto: Toronto Transit Commission.