Mining the Networks of Supply and Demand of Public Transport for Overload and Waste of Resources

Carlos Caminha, Vasco Furtado, Vládia Pinheiro e Caio Ponte
Programa de Pós-graduação em Informática Aplicada (PPGIA)
Universidade de Fortaleza (Unifor)
Brazil

Abstract—We propose here a methodology to help to understand the shortcomings of public transportation in a city via the mining of complex networks representing the supply and demand of public transport. We show how to build these networks based upon data on smart card use in buses via the application of algorithms that estimate an OD and reconstruct the complete itinerary of the passengers. The overlapping of the two networks sheds light in potential overload and waste in the offer of resources that can be mitigated with strategies for balancing supply and demand.

Keywords — Mobility, Complex Networks, Data Mining.

I. INTRODUCTION

In the era of digital information the biggest challenge is to, from diverse and voluminous data, create means to generate knowledge and apply it with effectiveness. One of the areas that best exemplify the context described above is the urban mobility. Sensors and digital media record daily information on tracks of people, which becomes a rich input for the carrying out of studies for understanding human mobility and its impact in public transportation, for instance.

In this article we describe our experience in a large Brazilian metropolis within the framework of its project for Smart City. From an origin-destination matrix estimated from data about the use of smart cards in buses, we created first a complex network that represents the demand of public transportation in the city. For doing this we have developed a heuristic algorithm to rebuild the complete itinerary of the passengers of the bus system. Another network, representing the supply of public transport considering the vehicles used, the routes followed and the itinerary of the routes with their bus stops, was produced.

The characterization of these networks and a strategy to overlap them, lead us to find out mobility patterns of persons using buses and how this generated knowledge to uncover problems of the supply and demand networks of public transport such as bottlenecks and/or waste of resources.

II. RELATED WORK

Recently a growing body of literature has been generated describing the complex networks modeling in context of air [3], rail [4] or urban bus [5] and bikes [6] transport. In general, these studies aim to characterize the networks using metrics such as the weight distribution, average path length and clustering coefficient. It is noteworthy mentioning the work of [7] who compared the public transport system in 22 towns in Poland. Among other characteristics, the authors showed that the degree and weight distribution follows a Power Law, in particular, a Small-World one that are hierarchically organized.

The work presented in [8] is particularly relevant in this context and is complementary to our. The paper presents an algorithm for detecting bottlenecks in a congested large-scale packet-switched network. The approach is based on estimating the per-link expected value of delay and turning this expected value to a per-link weight.

Regarding the ODM estimation, traditional methods are through large-scale sampled surveys, conducted once in every 1-2 decade. Two main disadvantages of these methods are the financial cost and obsolescence. Alternatively, models using the traffic count on a set of links have been proposed. The accuracy of these models depends on estimation model used, the input data errors and on the set of links with collected traffic counts [9]. In this work we are interested in static ODM for long-time transportation planning. The Entropy Maximization (EM) and Information Minimization (IM) models, and statistical approaches like Maximum Likelihood (ML), Generalized Least Squares (GLS) and Bayesian Inference (BI) have been mostly adopted. However, [9] highlight that few realistic approaches focused on large size network applications. More recently, Artificial Neural Networks for ODM estimation was applied on hypothetical networks with several constraints [10], and more studies still need to be carried out for planning and designing purpose for large size networks.
III. REPRESENTING THE PUBLIC TRANSPORT NETWORK THROUGH A COMPLEX NETWORK

A. Data about Public Transport

For the realization of this research five datasets were used related to network of public transport of Fortaleza in Brazil. The data refer to March 2015 and comprise the following: bus stops; routes; Terminals; bus GPS; and ticket validation. The bus stops are places where passengers go on and off the vehicles. In all, this data set has 4783 geo-referenced records. Bus routes have 359 records, which have information about the route direction and its itinerary. Bus terminals, the smallest of the data sets, as they possess only seven records, a terminal is a place designed to provide a better service in the user transshipment, various bus routes pass through terminals. The bus GPS dataset data has around 104 million records. The buses register their location every thirty seconds. In total 2034 buses circulated in Fortaleza in March 2015.

The user of bus in Fortaleza can use a smart card as a ticket for validating a trip. With the smart card it is possible to make connections in any point of the city as long as it is within the period of two hours since the last ticket validation. A validation occurs when the user passes his or her smart card in the bus or in a terminal validator.

The dataset of ticket validation refers to the month of March and has about 29 million recordings made in bus or directly at the validator of a bus terminal. The basic data are the user’s id, data and time of the validation, bus route identification, terminal identification (in case of validation in a terminal validator), and vehicle identification.

![Figure 1. Ticket validations in March](image)

Figure 1 illustrates the distribution of validations made in the days of March 2015. There is a pattern for days of the week, another for Saturdays (7th, 14th, 21st, 28th) and another for Sundays (1st, 8th, 15th, 22nd) and public holidays (the 19th and 25th). Practically all the days of the week that were not holidays there were approximately 1.1 million validations made, the only exceptions were the days 5 and 6, while the first there were 2.3 million validations and the second had only 838. The administrator of the database believes that there must have been some errors generating the file, for this reason, these data will be disregarded in our analysis.

B. Generating the Origin-Destination Matrix

To perform the origins estimates and the user’s destinations, the data already detailed previously were used. The problem that we faced was that the raw data did not allow saying precisely the bus stop where the passenger got on the bus, because they contain the time that a given user validated his ticket on the bus ratchet. We had first to estimate the place where this validation was done. For this reason, we retrieved in the bus GPS dataset the latitude and the longitude where the bus was at that validation time. However, this geographical coordinate is not necessarily the origin of the passenger, since he/she may not have done the validation of the ticket at the time that got on the bus. He could, for example, have remained at the rear of the vehicle up to the moment in which the bus approached from its destination. This problem can be minimized as we detail further.

With regard to the destination of the user unfortunately there is no precise record that indicates where the passenger got off. However, in literature, we found some heuristics that aim at supporting the estimate of that point of going off the bus with good reliability. The first is that, in their last journey in the day, the majority of people tend to return to the point where they set off [11, 12]. That is, in a day where the passenger has made only two trips, for example, the location of the first validation corresponds to the origin and the location of the second validation has good probability of being the destination. With this you can assume with certain reliability that the pair origin and destination (OD pair) of this user in question corresponds to the ticket validation coordinates.

The OD pairs may still be generalized to the cases where the passengers take more than two trips in a day. If the user has made three trips in a day, you can assume that the first trip corresponds to his first origin at the day, O1, the second trip has at its validation point at the first destination, D1, and the second origin of the day, O2. The third journey, in its turn, has in its Single ticket validation location at the second destination at the day, D2, and the third the origin, O3. Closing the triangle of OD pairs, it is assumed that the user will terminate their paths of the day returning to the point of first origin in the day, which in this case would also be the destination, D3, the user. The OD pairs in this example would then be \{O1, D1\}, \{O2, D2\} and \{O3, D3\} where in terms of geographical coordinates D1 = O2, D2 = O3 and O1 = D3. This example exposed in this graph may be associated with the case of a citizen who, for example, leaves the house in the morning toward the location where he or she works, then in another shift, goes to the university and then back home at the end of their daily routine.

In [11, 12] different strategies have been proposed for the validation of origin and destination. One of them says that the behavior of a recurring user must be checked. In days equal in the week, it is expected that the user leaves the same origins to the same destinations, for example, if in almost every
Monday of a month, a user "X" leaves a point P1 for a P2 and then to a point P3, it should be assumed that this is the default for moving the user "X" on Mondays. Then if in few Mondays it is recorded that he made only the route P1 to P3, it must be verified if these origins and destinations are correct. In this work it is verified if the route taken by the user from P1 arrives in P3. If not reached, it is verified if the same arrives in P2, and if he or she does, then it is added to the user analyzed the point P2 as an intermediary point in the itinerary. In case it is not possible to arrive in P2 from the bus route of P1, the OD is removed. This user scenario described usually occurs because possibly the user made part of the path that day by another modal such as bicycle, train or subway.

As it has already been mentioned, the point where the user made the ticket validation may not necessarily be the origin of the same, because this user may not have validated his or her ticket at the time that he or she got on the bus. The literature suggests [11,12] that to give more precision to these points, it must also be adopted a strategy to analyze the behavior of a recurring user. A person normally sets off from the same origins, but may have validated the ticket in different points of the bus route (capacity of the vehicle can motivate this), then it must be adjusted all the origins of the user in question to the first point of the validation found in the days analyzed. This is probably the closest point of origin of the user.

Using the strategies outlined in this section, the OD pairs have been generated of the bus network of Fortaleza for a week in the month of March 2015. The chosen period was from 11th (Wednesday) to 17th (Tuesday). This week has been chosen due to the absence of holidays and also due to the fact that the same does not possess the inconsistencies mentioned in section III of this work. In total, around 1.7 million OD pairs have been generated in this period.

C. Statistical validation of the representativeness of the OD sample spatial

Typically, ODM are constructed via surveys with a sample of the users. It is expected that the samples guarantee, with a certain degree of reliability, that the ODs estimated are correct. A major challenge is to ensure that the sample is not biased spatially, for example, that the embarkings into the bus used in estimates are not proportional (considering each bus stop) to the total of embarkings.

To spatially validate the sample we have generated we have checked the correlation between the embarkings used to estimate the OD pairs and the total of embarkings in the bus stops in Fortaleza. From the equation \( y = Y \cdot x^\beta \), where \( x \) represents the total of embarkings at the bus stops, \( y \) the total of embarkings used at the estimates and \( Y \) is a normalization constant, we say that a linear or isometric relationship, as evidenced by the value of the exponent \( \beta \) close to one, indicates that, proportionally, the embarkings used in estimates are equivalent to the total of embarkings in bus stops, noting that the samples were extracted in random order.

Figure 2 shows the correlation between the sample of the embarkings used in the estimation of origins and destinations and the total of bus embarkings recorded in the period studied. The red line represents the regression made about the data, it was estimated the exponent \( \beta = 0.95 \pm 0.02 \) and \( R^2 = 0.92 \). Also the confidence intervals of the correlation were estimated for the data distribution \( \{X_i, Y_i\} \). It was applied the method Nadaraya-Watson [13] to build the kernel smoother function,

\[
m_n(x) = \frac{\sum_{i=1}^{N} K_h(x - X_i)Y_i}{\sum_{i=1}^{N} K_h(x - X_i)},
\]

where \( n \) is the number of points of distribution and \( K_h(x - X_i) \) is a Gaussian kernel function formally defined as \( K_h(x - X_i) = \exp \left( \frac{(x - X_i)^2}{2h^2} \right) \), where \( h \) is the estimate of bandwidth by minimum squares using cross-validation [13,14]. We compute the 95% confidence interval (CI) 500 random bootstrapping samples with replacement.

![Figure 2](image)

**Figure 2.** The isometric relationship between the sample and the total of embarkings.

IV. NETWORKS: BUILD AND CHARACTERIZATION

A. Supply Transport Network

The transport infrastructure network of buses of a city is determinant in the movement of persons making the same reach their places of desire for performing daily activities. In the specific case of bus network, the nodes are bus stops and the bus routes that lead people between these bus stops are the edges. The itinerary of a route is the set of bus stops visited during a route. It is said that a route finished its itinerary when the bus reaches its last stop.

We can thus model a network as a directed graph, \( G(V, E) \), with vertices, \( V (\subseteq V) \), representing the bus stops and edges, \( E (\subseteq E) \), supplies between two bus stops. Formally the weight of the edge, \( w_{v_i \rightarrow v_j} \), represents the bus supply between two bus stops \( v_i \) and \( v_j \) (\( \in V \)). This supply is then calculated from the summing up of the bus routes weight, \( w_{l_{ij}} \), that pass by two bus stops. Formally \( w_{v_i \rightarrow v_j} = \sum_{l=1}^{N} w_{l_{ij}} \), where \( N \) is the total number of routes that visit, in sequence, \( v_i \) and \( v_j \). In
its turn, the weight of the bus routes, is calculated from the product of the quantity of vehicles, $V_i$, by the number of times that a vehicle of this route finishes its itinerary in a day, $C_i$. That is, $w_{Li} = V_iC_i$. The Figure 3 illustrate the generated supply graph.

![Figure 3. Supply graph from Fortaleza city.](image)

The concept of "completing the itinerary", done by a route, is fundamental in this calculation because the itineraries can have different sizes, a route that has more vehicles than another not necessarily has a higher supply. Said another way, the greater the weight of $w_{Li}$, the more the vehicles of $L_i$ pass by their itinerary’s bus stops.

**B. Demand Transport Network from the passenger route**

From the estimated OD matrix, we developed an algorithm to reconstruct the full path of an user in the bus network. The simplest case of route reconstruction is when the adopted route by the passenger takes him from the origin bus stop to the estimated destination. The case where there is no direct way to reach the destination by following only one route requires other heuristics. It is assumed in this case that the passenger made a connection with another line somewhere. The first problem in this situation is to identify this other line. This is trivial for cases where the connection was registered with the passage of the card in another bus indicating that it is the continuation of the route. The problem is that sometimes the connections are made at terminals where it is not necessary to make the card validation again. The passengers, who make connections at terminals, ascend the bus from the front door and do not need any card validation. One must then estimate in which terminal this connection has been made.

In these cases the aim is to identify the terminal(s) that are visited by both lines: leaving the source and arriving at the destination. That is, if there is a terminal in common on the route of these lines, it is assumed then that the transfer was made there. As a result, the passenger path becomes the route from source to terminal and from terminal to destination. In total, more than 1.2 million full path have been generated.

The demand network contains all the passenger paths that have been reconstructed. The path of a passenger is represented by a bus stop sequence. Formally it is a directed graph, $G (V, E)$ with vertices $v (\in V)$, representing the bus stops and edges $(\in E)$ representing the flow of people. The weight of the edge, $w_{vj}$, totals the amount of people who passe between $v_i$ and $v_j$.

Figure 4 illustrates the building process of this network. Given an user $U_1$ who makes the path containing the bus stop sequence $[v_3, v_4, v_5, v_6]$ and another user $U_2$, with path $[v_1, v_2, v_4, v_5, v_7, v_6]$, then the weight of the edge between $v_4$ and $v_5$ is incremented.

![Figure 4. Example of the network construction.](image)

**C. Networks Characterization**

In this section we describe the characterization process of supply and demand Fortaleza bus system. To allow comparisons, edges of the weights of the two networks were normalized. In the demand network, for each edge, $e_d$, the weight $w_d$, describing the demand of passengers between two bus stops, was normalized by the ratio $w_d/w_{d\text{max}}$ where $w_{d\text{max}}$ is the largest weight registered on the network. For the supply network the same strategy was used for computing the weight that represents the amount of available buses. Note that both networks have the same degree distribution, i.e., for each edge in the supply network there is a corresponding edge on demand network, they differ only with respect to the weights of the edges.

Figure 5 illustrates the weight distribution for the two networks. In Figure 5 (a) it is shown a Power Law $[25, 26]$ with exponente $\alpha = -2.90$ for the Supply network. This result suggest that this network was projected for offering a high number of resources in few places and for supplying few resources in several places. Similarly, the Power Law found for the demand network and depicted in Figure 5 (b) with...
exponente $\alpha = -1.97$ indica que hay pocos lugares con alta demanda y varios otros lugares con baja demanda.

Figure 5. Weight Distributions of supply and demand networks. In (a), the green line shows the regression applied to the data in (b) the regression is represented by the orange line. The blue dashed lines represent confidence intervals estimated with Nadaraya Watson method. The distributions were generated using 50 bins logarithmic.

Después de que las distribuciones presenten similitud en términos de escala, esto no necesariamente implica que los sistemas estén balanceados. La primera pista de imbalances es la diferencia entre los exponentes, el volumen de concentración en pocos lugares es mucho más intenso en la red de suministro. Figura 6 muestra la diferencia de pendiente de las estimaciones de regresión.

Aún en las distribuciones de pesos de las redes, en la Figura 7 se ilustran las distribuciones acumulativas de estos datos. La curva naranja muestra que la red de demanda acumula más aristas con pesos pequeños que la red de suministro.

Figura 6. Comparación de las escalas de las leyes del poder encontradas. La línea verde muestra la regresión estimada en la red de suministro y la línea naranja muestra la regresión estimada de la distribución de pesos de las aristas de demanda.

Figura 7. Distribuciones acumulativas de los pesos de las aristas en ambos sistemas estudiados. La curva naranja representa la distribución acumulativa de los pesos de las aristas de demanda, la curva verde representa la correspondiente distribución acumulativa de la oferta.

Hemos investigado también si los sistemas (componente grande) podrían ser divididos en sub redes con alta coeficiente de agrupamiento interno y baja conectividad con componentes externos. Esto podría indicar la existencia de áreas donde el suministro de bus es privilegiado y también de puntos críticos (lacunas).

We have investigated also if the networks (giant component) could be divided into sub nets with high internal clustering coefficient and low connectivity with external components. This might indicate the existence of areas where the bus supply is privileged and also bottlenecks (weak links).
between these areas. It was used the algorithm [15] to detect the communities. This algorithm makes use of a heuristic method based on the optimization of modularity of graphs. The modularity of a set of nodes is measured by an actual value between 0 and 1, being computed by the ratio between the quantity of edges that connect the elements of the set among themselves by the total of the edges of the set of nodes [27-31]. The algorithm can be divided in two phases that are repeated iteratively. In the first phase it is assigned a different community for each network node. Thus, in this initial split there are many communities as nodes. For each node \( i \) are considered the neighbors of \( i \) and \( j \) is evaluated the gain of modularity that occurs when moving \( i \) of their community to the community of \( j \). The node \( i \) is then placed in the Community for which this gain is maximum, but only if this gain is positive. If no positive gain is possible, \( i \) remains in their original community. This process is applied to all nodes repeatedly until no improvement can be achieved and the first phase is thus completed. The second phase of the algorithm consists of the construction of a new network whose nodes are the communities found during the first phase. To do this, the weights of the connections between the new nodes are given by the sum of the weight of the connections between the nodes in the two corresponding communities. The connections between the nodes of the same community are represented by self-loops for this node of the new network. The phases are repeated until the maximum global modularity is found.

**Figure 8.** Comparison between the modularity of the two networks. The green line shows the modularity of the supply network according to the network is divided in more communities. The line orange illustrates the modularity behavior of the demand.

Figure 8 illustrates the increased modularity when the network is divided into more communities. The modularity of the supply network is greater than the demand network in virtually all configurations. This result reveals that the supply network is designed to meet shorter movements than the fact that people need, because the modularity found in this network indicates that there are more strong links intra-community than those found in demand network, possibly because the users of the network travels over longer stretches, increasing the weight of the edges in these locals and preventing community detection algorithms find modular configurations such as those found in the supply network.

In balanced networks, where supply grows in direct proportion to the growth of demand, it is expected that there is an isometric relationship between supply and demand volumes. In Figure 9 we measure the correlation between the weights of edges of supply and demand networks. Low correlation, \( R^2 = 0.53 \), reveals an imbalance between supply and demand, and that the value of \( \beta = 1.24 \) reveals an allometric relationship [20-24], different from the expected isometric relationship. Above the red line (linear regression) are the edges or parts of the network, where demand is proportionally greater than the supply. However, it draws attention the highest concentration of dots below the regression line, indicating that there are parts of the city in where resources offered are, proportionally, greater than the existing demand in the region. This signals that one can best meet the demand of the network without the need to increase the operating cost of the same, just trying to balance supply in some parts of the city.

**Figure 9.** Correlation between the volume of supply and volume of demand on all edges of the network. Each black dot represents an edge. The x-axis represents the weight supply edge and the y-axis the weight of demand edge. The red line is the regression applied to the data. The blue dashed lines illustrate the estimated confidence intervals with Nadaraya Watson method [13,14]. The data are shown in logarithmic scale.

Spatially, the city places (segments between two bus stops) where the network is potentially overloaded by demand can be seen in Figure 10. We have them created a third network, where each edge \( w_{sd} \) was assigned \( \frac{w_s}{w_{s_{max}} - w_d/w_{d_{max}}} \). Subtracting the edges of weights generates new weights with values between 1 and -1, where negative values
indicate overload on the edge. 14.88% of the edges of Fortaleza bus network have \( w_{ed} < 0 \).

**Figure 10.** Overlay of the demand for network supply network. The edges in bold represent those where demand is proportionally greater than the supply. The red dots highlight the locations of the 7 bus terminals of the city.

V. CONCLUSION

This work explored the supply and demand for bus networks in a case study with a large Brazilian metropolis. Models and metrics of complex networks have been used to understand the dynamics of urban mobility. Our main contribution is THE propose a model based on complex networks with potential to discover bottlenecks and resource waste in bus networks in big cities.

Secondary contributions of this work are, first, the reconstruction and statistical validation of the origins and destinations of users from public data bus GPS and, second, the reconstruction of paths of users within the bus system making use of heuristics based on the behavior of these users.

REFERENCES

[1] Wang, W., Attanucci, J., Wilson, N. Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems. Journal of Public Transportation, 14 (4): 131-150. 2011.
[2] De Solà-Morales, M.: "The strategy of urban acupuncture." Structure Fabric and Topography Conference, Nanjing University. 2004.
[3] Exploring the network structure and nodal centrality of China’s air transport network: A complex network approach. Journal of Transport Geography. 2011, Pages 712–721.
[4] Lenomard, M. Tugores, A. Colet, P. Ramasco, J. Tweets on the Road. PLOSone, v.9, issue 8, 2014.
[5] Munizaga, M. Palma, C. Fischer, D. Estimation of a Disaggregate Multimodal Public Transport OD Matrix from Passive Smart Card Data from Santiago, Chile. TRB 2011 Annual Meeting, Washington, 2011.
[6] Hamon, R. Borgnat, P. Flandrin, P. Robardet, C. Networks as signals, with an application to bike sharing system. GlobalSIP 2013, Austin, Texas, USA, Dec. 2013.
[7] Sienkiewicz, J. and Holyst, J Statistical analysis of 22 public transport nets in Poland. 2005.
[8] Elmasry, G. F. and McCann, C. J., Bottleneck discovery in large-scale networks based on the expected value of per-hop delay, Military Communications Conference, 2003. MILCOM ’03. 2003 IEEE, pp. 405-410 Vol.1. 2003.
[9] Sharminda Ber, K.V. Krishna Rao. Estimation of origin-destination matrix from traffic counts: the state of the art. European Transport, n.49, 3-23, 2011.
[10] Remya, K.P, Mathew, S. OD matrix estimation from Link Counts Using Artificial Neural Network. International Journal of Scientific & Engineering Research. V.4, Issue 5, p.293-296. 2013.
[11] Gordillo, F. The Value of Automated Fare Collection Data for Transit Planning: An Example of Rail Transit OD Matrix Estimation. Massachusetts Institute of Technology. [S.l.]. 2006.
[12] Hua-Ling, R. Origin-Destination Demands Estimation in Congested Dynamic Transit Networks. International Conference on Management Science & Engineering (14th), Harbin, Agosto 2007.
[13] Nadaraya, E. On estimating regression. Theor. Probab. Appl. 9, 141–142. 1964.
[14] Watson, G. S. Smooth regression analysis. Sankhya Ser. A 26, 359–372. 1964.
[15] Blondel, D. Guillaume, J. Lambiotte, R. Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, Issue 10, pp. 10008, 12 pp. 2008.
[16] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 8271-8276, 2002.
[17] Wilkinson, D. and Huberman, B. A method for finding communities of related genes. In Proc. Natl. Acad. Sci. USA, 2004.
[18] Harary, F. Graph Theory. Reading, MA: Addison-Wesley, p. 35. 1994.
[19] Freeman, L. A set of measures of centrality based on betweenness. Sociometry. 40, 35-41. 1977.
[20] Oliveira, E. A., Andrade, J. S. and Makse, H. A. Large cities are less green. Scientific Reports, 4, 4235, 2014.
[21] Bettencourt, L., Lobo, J., Helbing, D., Kühnert, and West, G. Growth, innovation, scaling, and the pace of life in cities, 2007.
[22] Piaget, H., Moreira, A., Batista, E., Makse, H. and Andrade, J., Statistical signs of social influence on suicides. Scientific reports, 2014.
[23] Chen, Y. An allometric scaling relation based on logistic growth of cities, 2013.
[24] Naroll, R., Bertalanfly, L. The principle of allometry in biology and social sciences. General Systems Yearbook, 1(part II): 76–89, 1956.
[25] Barabasi, R., Albert, R. Jeong, H. Scale-free characteristics of random networks:the topology of the world-wide web. Physica A 281, 69-77, 2000.
[26] Clauset, A. Shalizi, R. Newman, M. Power-law Distributions in Empirical Data. SIAM Review 51, 661-70, 2009.

[27] M. Girvan and M. E. J. Newman. Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 8271-8276, 2002.

[28] Wilkinson, D. and Huberman, B. A method for finding communities of related genes. In Proc. Natl. Acad. Sci. USA, 2004.

[29] Tyler, J., Wilkinson, D. and Huberman, B. Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. Communities and Technologies. eds. Kluwer Academic, 2003.

[30] Girvan, M. and Newman, J. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826, 2002.

[31] Wu, F.; Huberman, B. Finding communities in linear time: a physics approach. The European Physical Journal B, Volume 38, Issue 2, March II 2004, pp.331-338, 2004.