BU-1630-M

Scaling laws for the movement of people between locations in a large city

G. Chowell[1,2], J. M. Hyman[2], S. Eubank[2], C. Castillo-Chavez[1,2]

[1] Los Alamos National Laboratory

Los Alamos, NM 87545, U.S.A.

[2] Department of Biological Statistics and Computational Biology, Cornell University,

Warren Hall, Ithaca, NY 14853-7801, U.S.A *

March 4, 2003

Abstract

Mathematical models and simulations of the movements of people in a large city can provide new insights into understanding and predicting dynamical systems that depend on the interactions between people. These applications range from optimizing traffic flow, slowing the spread of infectious diseases to predicting the change in cell phone usage in a disaster. We analyzed the simulated movements of 1.6 million individuals in a computer

*Los Alamos Unclassified Report LA-UR-02-6658.

(agent-based) model for Portland, OR. The dynamics (movement) of individuals are modeled as discrete events moving people between nodes on a directed weighted graph. The 181,205 nodes represent physical locations such as buildings and the connecting edges between the nodes have weights based on the daily traffic (movement of individuals) between the locations. The resulting weighted graph is scale-free and has scaling laws consistent with an underlying hierarchical structure. The number of edges leaving a node distribution (out-degree), the edge weights (out-traffic), and the edge-weights per location (total out-traffic) are power law distributions. This distribution is observed in subgraphs based on work, school, and social/recreational activities, which is consistent with an underlying hierarhical structure. We identified a linear correlation between the out-degree and the total out-traffic distributions. The short average distance between nodes on this sparsely connected network and the significant levels of clustering observed make it a small world [3][4]. The predictions of models on a network depend on the behavior power laws and hierarchical structure of the mobility of people in a city. Therefore, these distributions should be preserved in modeling for the spread of disease and other simulations where social dynamics and mixing of the population is important.

1 Introduction

Patterns have been detected in networks describing scientific collaboration [6][7][9], cellular networks [11][12], the Internet [14], and the *World Wide Web* [10][13]. These networks exhibit the "small world effect," [3][5] where the average number of edges needed to connect *any* pair of nodes is small and the network exhibits high levels of clustering, a characteristic absent in

random networks [1].

A network node has degree k if it is connected to k other nodes. The local connectivity of a network depends on the proportion of the nodes with degree k, $P(k)$, as a function of k. The connectivity of many networks based on real world data have a power-law distribution, $P(k) \propto k^{-\gamma}$, where the exponent γ characterizes the underlying scaling of the network.
The accuracy of computer generated networks used to simulate these systems depends on how faithfully they preserve the structure of real world data. Barabási and Albert (BA) introduced a model capable of generating networks with a power-law connectivity distribution with $\gamma = 3$. The BA algorithm generates networks where nodes connect, with higher probability, to nodes that have a accumulated higher number of connections. This algorithm stochastically generates networks with connectivity distributions $P(k) \propto k^{-\gamma}$ (in the appropriate scale) where k gives the incidence (number of connections) per node. if $\gamma < 2$ the average connectivity $\bar{k}$ is infinite. The variance is infinite whenever $\gamma < 3$.

We generate a directed graph for the simulated movement of 1.6 million individuals *in* or *out* of 181,205 locations in Portland, OR. The 181,205 nodes represent locations in the city and the edges connections between nodes. The edges are weighted by the daily traffic (movement of individuals) *in* or *out* of the locations. The statistical analysis of the network topology revealed that it is a small world with power-law decay in the out-degree distribution of locations (nodes) at different time scales. The out-traffic (weight of the full network) and the total out-traffic (total weight of the out edges per node) distributions are also power laws. We also observe that

the joint distribution of the out-degree and total out-traffic distributions decays linearly.

2 Location-based network

From a "typical" realization by the Transportation Analysis Simulation System (TRANSIMS) of the simulated dynamics of 1.6 million individuals in the city of Portland, we create a directed network where the nodes represent locations (i.e. buildings, households, schools, etc.) and the directed edges (between the nodes) model the movement (traffic) of individuals between locations (nodes). The mobility matrix $W = (w_{ij})$ is the nonsymmetric square matrix of weights assigned to all the directed edges in the network. That is, w_{ij} is the weight of the edge that goes from node i to node j (direction matters). We define $w_{ij} = 0$ when there is no directed edge connecting from node i to node j for the location-based network of the city of Portland.

TRANSIMS [18] is an agent-based simulation for the movement of individuals in a virtual region or city using empirical distributions constructed from survey data. A detailed representation of the region is created, mobility information for each individual is determined (in the case of Portland) via surveys and the transportation infrastructure is considered. TRANSIMS then simulates the movement of individuals in the virtual region through the transportation network via mass transportation or by car. The simulation is tunned to agree with movement data, obtained from transportation planning surveys of detailed information on people's movement (daily trips). The data also reflects the activity types (see Figure 1), origins, destinations, routes, timing, and forms of transportation used. The TRANSIMS model calculates the simulated movements of 1.6 million individuals in a typical day [18]. Similarly generated data

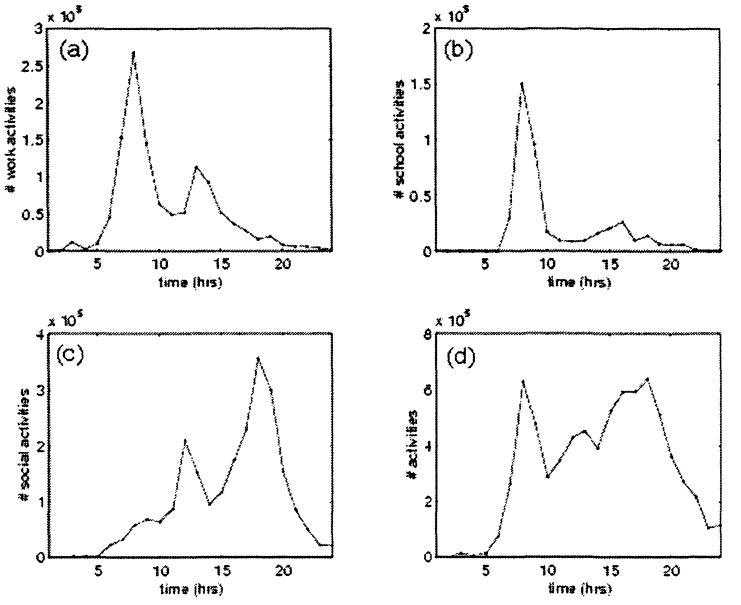

Figure 1: Activity is the movement of an individual to the location where the activity will be carried out. (a)work activities, (b)school activities, (d)social activities, and (d)total number of individual activities as a function of time (hours) of a 'typical' day in the city of Portland, OR.

from the cities of Albuquerque and Dallas will be used in the future and compared with the observations in this paper. The simulated Portland data set includes the time at which each individual leaves a location and arrives at another. From this, we can calculate the average number of people at each location and the traffic between any two locations on a typical day. Table 1 shows a sample of a Portland activity file generated by TRANSIMS.

TRANSIMS estimates of the social network in a large urban area are based on the assumption that the transportation infrastructure constrains people's choices about what activities to perform and where to perform them[18]. A synthetic population is endowed with demograph-

ics matching the joint distributions given in census data. Observations are made on the daily activity patterns of several thousand households (survey data). These patterns are used as templates and associated with synthetic households with similar demographics. The locations at which activities are carried out are estimated taking into account observed land use patterns, travel times, and dollar costs of transportation alternatives. Estimated locations are fed into a routing algorithm to find minimum cost paths through the transportation infrastructure consistent with constraints on mode choice [22, 23, 24]. The simulation resolves distances down to 7.5 meters and times down to 1 second. It provides an updated estimate of time-dependent travel times for each edge in the network, including the effects of congestion, to the Router and location estimation algorithms, which produce new plans. The resulting traffic patterns compare well to observed traffic. The entire process estimates the demand on a transportation network from census data, land use data, and activity surveys. It can thus be applied to assess the effects of hypothetical changes such as building new infrastructure or changing downtown parking prices. Methods based on observing demand cannot handle such situations, since they make no claim to understanding what generates the demand.

Though social network estimates are not strictly speaking part of TRANSIMS, the level of detail in the simulation allows us to aggregate the movement of individuals at multiple time and space scales to compare with real world observations and search for hierarchical scaling. Until recently, it has been difficult to obtain these estimates. Typically, certain classes of random graphs (scale-free networks [8], small-world networks [5], Erdos-Renyi random graphs [2][1], or ad-hoc networks [?]) have been postulated as good representatives. Alternatively, models have

been restricted to small scales where direct observation is feasible [21].

Table 1. Sample TRANSIMS's activity file. Hence, one sees that person 115 arrives for a social recreational activity at location 33005 at 19.25 o'clock. This person then departs at 21.00 o'clock.

Person ID	Location ID	Arrival time(hrs)	Departure time(hrs)	Activity type
115	4225	0.0000	7.00	home
115	49296	8.00	11.00	work
115	21677	11.2	13.00	work
115	49296	13.2	17.00	work
115	4225	18.00	19.00	home
115	33005	19.25	21.00	social/rec
115	4225	21.3	7.00	home
220	8200	0.0000	8.50	home
220	10917	9.00	14.00	school
220	8200	14.5	18.00	home
220	3480	18.2	20.00	soc/rec
220	8200	20.3	8.6	home

3 Power law distributions on the Portland network

We calculate the statistical properties of the location-based network of the city from the mobility data generated by the TRANSIMS simulation for a typical day. The Portland simulation has

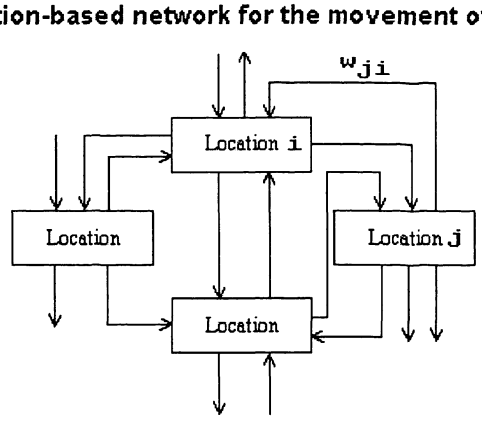

Figure 2: The nodes represent locations connected via directed edges based on the traffic or movement of individuals between the locations. The weights (w_{ij}) represent the daily traffic from location i to location j.

power law distributions for the out-degree, the out-traffic (edge weights) and the total out-traffic(edge-weights per node). The analysis confirms that there are only a few degrees of sparation between any two nodes [3] and the power law decay is preserved in many subnetworks. This is consistent with an underlying hierarchical topology.

The average out-degree, $\bar{k}$, is defined as the individual node out-degrees of the whole network divided by the total number of nodes,

$$\bar{k} = \sum_{i=1}^{n} k_i/n$$

.

Here k_i is the degree for node i and n is the total number of nodes in the network. The clustering coefficient, C, quantifies the extent to which neighbors of a node are also neighbors of

each other [4]. The clustering coefficient of node i, C_i, is given by

$$C_i = |E(\Gamma_i)| / \binom{k_i}{2}$$

where $|E(\Gamma_i)|$ is the number of edges in the neighborhood of i (edges connecting the neighbors of i not including i itself) and $\binom{k_i}{2}$ is the maximal number of edges that could be drawn among the k_i neighbors of node i. The clustering coefficient of the whole network is the average of the individual clustering coefficients C_i's, that is, $C = \sum_{i=1}^{n} C_i/n$.

The average distance between nodes, L_i, is defined as the median of the means of the shortest path lengths connecting a vertex $i \in V(G)$ to all other vertices [4]. $\bar{L}$ is the median of the L_i's.

The out-degree distribution, $P_d(k)$, gives the probability that a randomly chosen node from the network has out-degree k. The out-traffic distribution, $P_o(u)$ gives the probability that a randomly chosen edge from the network has weight u. That is, $P_o(u)$ is the probability distribution of the matrix W of *weights*. The total out-traffic distribution, $P_t(v)$ denotes the probability that a randomly selected node from the network has a total out-traffic v (sum of its edge-weights). That is, $P_t(u)$ is the probability distribution of w_i given by:

$$w_i = \sum_j w_{i,j}$$

Finally, the joint distribution of the degree and the total out-traffic distributions, $F(k, v)$ gives the probability that a randomly chosen node from the network has out-degree k and total out-traffic v.

Table 2. Computed statistics of the location-based network of the city of Portland.

Statistic	Value
Total nodes (n)	181205
Total directed edges (m)	5416005
Average degree ($\bar{k}$)	29.88
Clustering coefficient (C)	0.042
Average distance between nodes (L)	3.38
Diameter (D)	9.0

Watts and Strogatz [3] compute the clustering coefficient C by listing all the neighbors of a node, counting the edges that link those neighbors, and dividing by the maximum number of edges that could be drawn from all the node's neighbors. The process is repeated on all the nodes and then averaged. For a typical random graph with $181,205$ nodes and an average degree $\bar{k} = 29.88$, this approach gives a clustering coefficient $C = 0.000166$. The clustering coefficient for our location-based network, ignoring edge directions, is $C = 0.0584$, which is

is roughly 350 times larger than the expected clustering coefficient for a random graph of the same size and average degree $\bar{k}$. This confirms that the location-based network for Portland is highly clustered. Highly clustered networks have also been found for other real-world systems [3]. For example, the electric power grid of western US has a clustering coefficient $C = 0.08$, about 160 times larger than the expected value for an equivalent random graph [4].

For the location-based network of the city of Portland, the average distance between two nodes is $L \approx 3.38$ and the largest of the shortest paths between locations (diameter of the graph) is $D = 9$. L and D were measured using a BFS (breadth first search) algorithm [25] on a randomly selected subgraph of size $90,000$ ($\approx 50\%$ the size of the whole network) ignoring the edge directions. The few degrees of separation between the nodes on this highly clustered network both contribute to the small world effect.

Many real-world networks exhibit properties that are consistent with an underlying hierarhical organization. Hierarchically organized networks have groups of nodes that are highly interconnected, but these groups have only a few edges or no edges connected outside of their group. Hierarchical structures of this type can be characterized by the clustering coefficient function $C(k)$ as a function of the degree k. For the network of movie actors, the semantic web, the *World Wide Web*, the Internet at the autonomous system level is the degree, and metabolic networks [19][20], the clustering coefficient scales as k^{-1}. The clustering coefficient as a function of the degree (ignoring edge directions) in the Portland network follows a similar scaling law on the whole network and on subnetworks constructed by type of activity (work, school and

social/recreational activities, see Figure 3). The clustering coefficient of the resulting subnetworks from work, school, and social activities are: 0.0571, 0.0557, and 0.0575, respectively. The largest clustering coefficient and closest to the overall clustering coefficient ($C = 0.0584$) correponds to the subnetwork constructed from social/recreational activities. It seems that the whole network, as well as the activity subnetworks, exhibit a hierarchical structure but the nature of such structure, as defined by the power law exponent, is not universal. This agrees with the theoretical observations in [20].

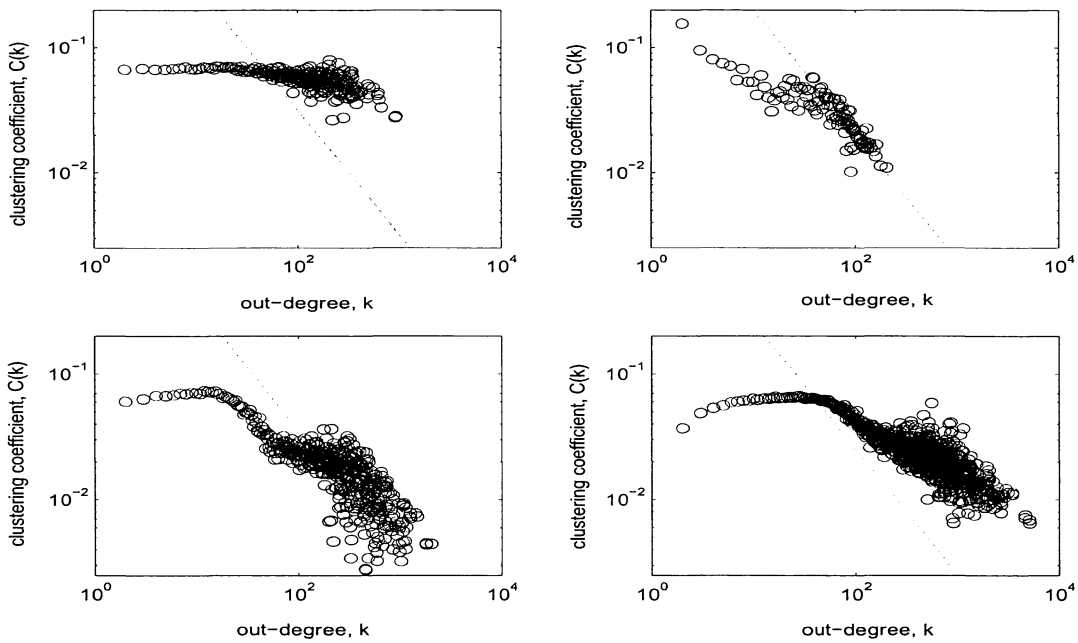

Figure 3: Log-log plots of the clustering coefficient as a function of the out-degree for (a)the location-based network constructed from work activities, (b)school activities, (c)social activities, and (d)all the activities. The dotted line has slope -1.

We computed the out-degree distribution of the network at three different time intervals: The morning from 6 a.m to 12 p.m.; the workday from 6 a.m. to 6 p.m.; and the full 24 hours. For the morning, the out-degree distribution has a tail that decays as a power law with $\gamma \simeq 3$, for the workday $\gamma \simeq 2.6$, and for the full day $\gamma \simeq 2.7$. The distribution of the out-degree data has two scaling regions. The number of locations is approximately constant for out-degree $k < 20$ and then decays as a power law for high degree nodes (Fig. 4). The clustering coefficient is approximately $C = 0.057$ for work activities but changes during the day as people engages in different activities. The daily average eliminates this time dependence and may not an appropriate measure of the network connectivity for events that only last a few hours.

The strength of the connections in our location-based network is measured by the traffic (flow of individuals) between locations in a "typical" day of the city of Portland. The log-log plot of the out-traffic distributions for three different periods of time (the distribution of W) in Fig. 5 exhibits power law decay with different exponents, $\gamma \simeq 3.56$ for the morning, $\gamma \simeq 3.74$ for the workday, and $\gamma \simeq 3.76$ for the full day. The out-traffic distribution is characterized in its entirety by a power law distribution for all values of the weight matrix $W = w_{ij}$. This is not the case for the out-degree distribution of the network (see Figure 4) where a power law fits well only for sufficiently large degree k.

The distribution of the w_i's is the distribution of the total out-traffic of locations. This distribution has a tail that decays as a power law with exponent $\gamma = 2.74$ (see Figure 6). This power-law exponent is almost identical to that of the out-degree distribution which may be expected since

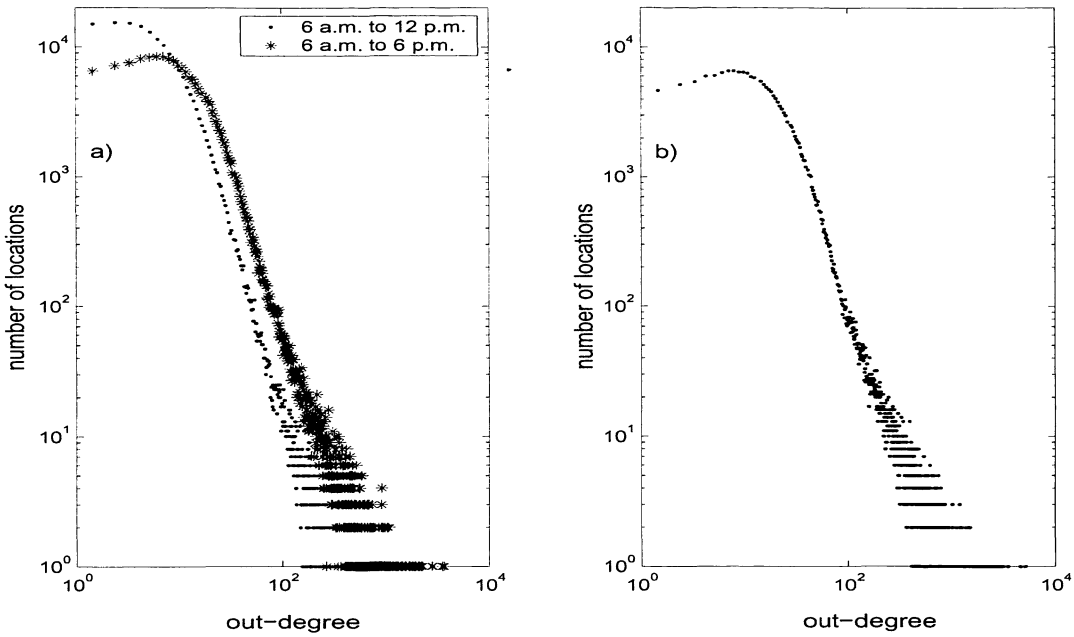

Figure 4: There are two distinct scaling regions for the number of nodes as a function of the out-degree of the nodes. There are approximately the same number of nodes with out-degree $k = 1, 2, ...10$ For $k > 0$ the number of nodes with a given out-degree decays as a power law $P(k) \propto k^{-\gamma}$ with $\gamma \simeq 3$ for the morning (6 a.m.-12 p.m.), $\gamma \simeq 2.6$ for the workday (6 a.m.-6 p.m.) and $\gamma \simeq 2.7$ for the full day.

the movement of individuals between locations contributes to both the out-degree and the total out-traffic. In the next section, we model the degree of correlation between the out-degree and the total out-traffic and we use them to gain some insight into their joint distribution.

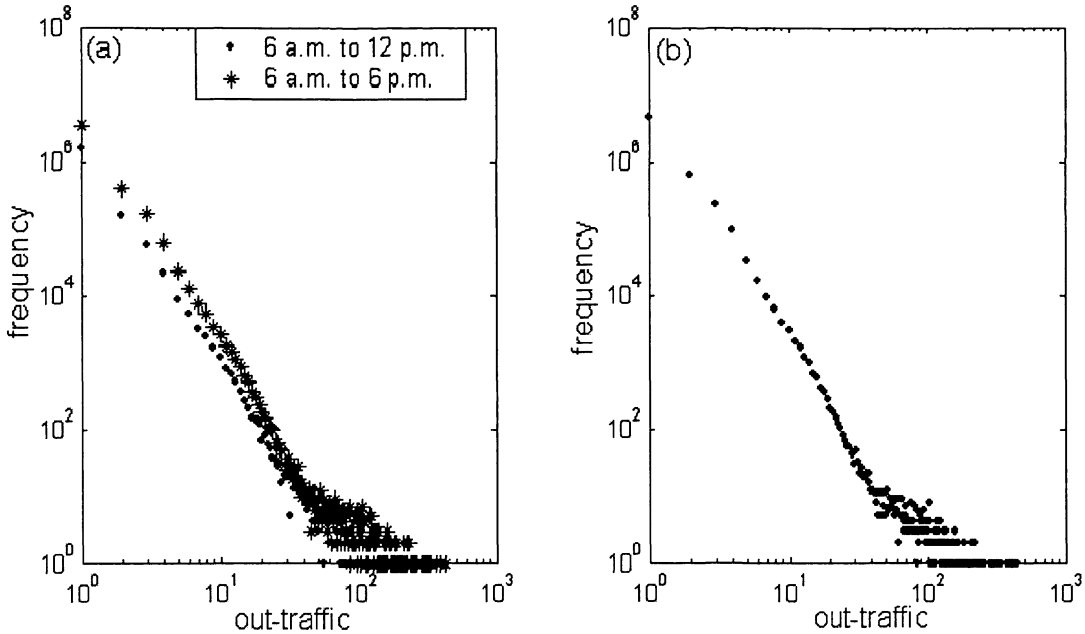

Figure 5: The out-traffic plots of the location-based network of the city of Portland for different time periods of the day: (a)For the first (6 a.m. to 12 p.m.) the power-law exponent $\gamma \approx 3.56$, for the second (6 a.m. to 6 p.m.) $\gamma \approx 3.74$, and for (b)the 24-hour period $\gamma \approx 3.76$. This power law distributions show high traffic in few connections while low traffic is observed in most connections.

4 Correlation between out-degree and total out-traffic

Out-degree k and total out-traffic v are highly correlated (Fig. 7) with a correlation coefficient $\rho = 0.94$ on a log-log scale with 95% of the nodes (locations) having out-degree and total out-traffic less than 100. That is, the density of their joint distribution $F(k, v)$ is highly concentrated near small values of the out-degree and total out-traffic. This joint distribution is characterized by a surface that decays linearly when the density is in log_e scale (see Figure 8).

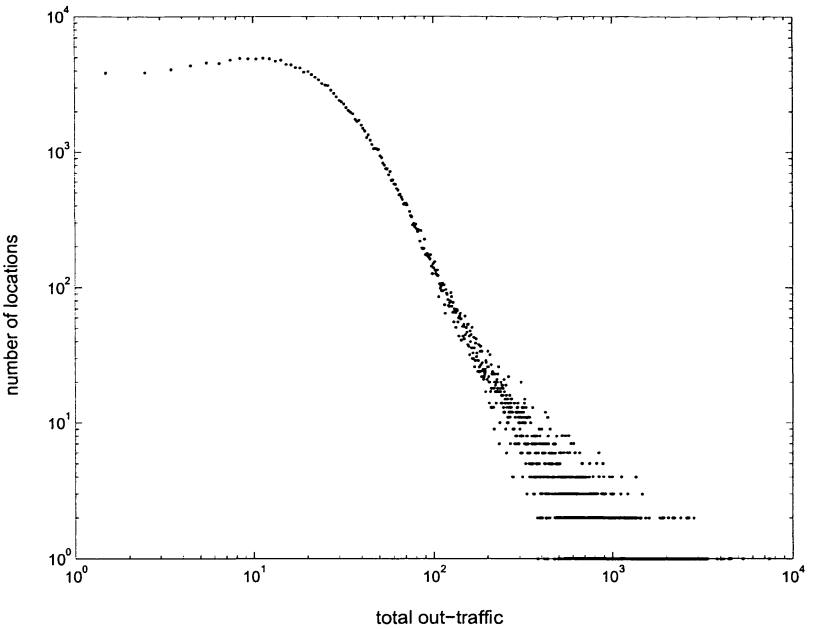

Figure 6: Distribution of the total traffic for the location-based network of the city of Portland has two scaling regions. There are approximately the same number of locations (nodes) with out-traffic=2,3,4,...30 people. The number of locations where more than 30 people leave each day decays as a power law with $\gamma \simeq 2.74$.

The degree of correlation between various network properties depend on the social dynamics of the population. Understanding the mechanisms behind the correlations will be key to creating high fidelity networks to model behavior changes in a population for simulations of special situations or events such as modeling disease transmission.

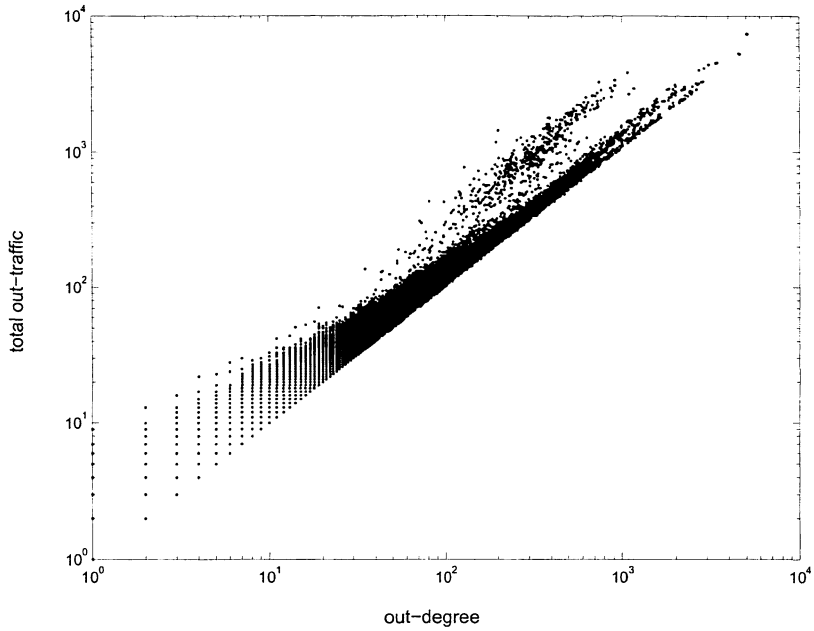

Figure 7: The out-degree and the total out-traffic are highly correlated with correlation coefficient $\rho = 0.94$ on a log-log scale. Most (95%) of the locations have fewer than 100 people leaving during the day.

5 Conclusions

We analyzed the movement of 1.6 million individuals in the city of Portland as a weighted directed graph where the nodes of the graph correspond to physical locations and the directed edges connecting the nodes are weighted by the number of people moving in and out of the locations during a typical day. The clustering coefficient, measuring the local connectedness of the graph, scales as k^{-1} (k is the degree of the node) for $k >?$. This is consistent with an underlying hierarhical structure where a few nodes get most of the action. The out-degree distribution network is relatively constant for small k and power law decay afterwards ($P(k) \propto k^{-\gamma}$). The

distribution of the daily total out-traffic between nodes is also flat for small k and exhibits power law decay afterwards. However, the distribution of the daily out-traffic of individuals between nodes scales as a power law even for small k.

The observed power law distribution in the out-traffic (edge weights) is in agreement with the theoretical analysis of Yook *et al.* [15]. They built weighted scale-free (WSF) dynamic networks and proved that the distribution of the total weight per node (total out-traffic in our network) is a power law. However, the weights in their WSF model decay exponentially.

The daily out-degree is highly correlated to the total out-traffic. This correlation is consistent with an underlying hierarhical structure where a few nodes have most of the connections and get most of the traffic (weight). We observed that the joint density of out-traffic is linearly correlated to the total out-traffic on a log-log scale. The joint probability distribution of the characteristics of networks can help guide in identifying the key mechanisms (hierarchy, etc.) that must be considered when generating networks for modeling the movement of people in a large city. Networks based on these mechanisms for the generation of graphs, will provide more "realistic" weghted networks that could be used in the study of various processes such as the spread of diseases in a large city.

6 Acknowledgements

This work was carried out as part of a summer research experience at Los Alamos National Laboratories (T-7 group). Gerardo Chowell is thankful to Pieter Swart, Leon Arriola, and

Albert László Barabási for interesting and helpful discussions.

References

[1] B. Bollobás, *Random Graphs*, Academic, London (1985).

[2] P. Erdos and A. Renyi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[3] D. J. Watts and S. H. Strogatz, *Collective Dynamics of Small-World Networks*, Nature, 363:202-204 (1998).

[4] D. J. Watts, *Small Worlds: The dynamics of networks between order and randomness*, Princeton University Press (1999).

[5] S.H. Strogatz, *Exploring Complex Networks*, Nature 410, 268-276 (2001).

[6] M. E. J. Newman, *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci. 98, pp. 404-409 (2001).

[7] M. E. J. Newman, *Who is the best connected scientist? A study of scientific coauthorship networks*, Physical Review E 64 (2001) 016131; Phys.Rev. E64 (2001) 016132.

[8] Albert-László Barabási, Réka Albert, Hawoong Jeong, *Mean-field theory for scale-free random networks*, PHYSICA A 272 (1999) 173-87.

[9] A.-L. Barabsi, H. Jeong, R. Ravasz, Z. Nda, T. Vicsek,and A. Schubert, *On the topology of the scientific collaboration networks*, Physica A 311, 590-614 (2002).

[10] R. Albert, H. Jeong, and A.-L. Barabási, *Diameter of the World Wide Web*, Nature 401, 130-131 (1999).

[11] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabási, *The large-scale organization of metabolic networks*, Nature 407, 651-654 (2000).

[12] H. Jeong, S. Mason, A.-L. Barabási, and Z.-N. Oltvai, *Lethality and centrality in protein networks*, Nature 411, 41-42 (2001).

[13] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew S. Tomkins, Eli UpfalProc (200). 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS.

[14] M. Faloutsos, P. Faloutsos, C. Faloutsos, *On Power-Law Relationships of the Internet topology*, SGCOMM (1999).

[15] S.H. Yook, H. Jeong, A.-L. Barabási and Y. Tu, *Weighted Evolving Networks*, Physical Review 86, 5835-5838 (2001).

[16] S. Eubank, *Episims Input and Output, Technical report*, Los Alamos National Laboratory (2001).

[17] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, *Classes of small-world networks*, Proc. Natl. Acad. Sci. (2000).

[18] C. L. B. . et al. TRANSIMS: Transportation Analysis Simulation System. Technical Report LA-UR-00-1725, Los Alamos National Laboratory, Unclassified Report, 2001.

[19] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai,and A.-L. Barabási, *Hierarchical Organization of Modularity in Metabolic Networks.* Science 297, 1551-1555 (2002).

[20] Erzsébet Ravasz and Albert-László Barabási, *Hierarchical organization in complex networks*, Physical Review E (in press).

[21] D. Peterson, L. Gatewood, Z. Zhuo, J. J. Yang, S. Seaholm, and E. Ackerman. Simulation of stochastic micropopulation models. *Computers in Biology and Medicine*, 23(3):199–210, 1993.

[22] C. Barrett, K. Bisset, R. Jacob, G. Konjevod, and M. Marathe. An Experimental Analysis of a Routing Algorithm for Realistic Transportation Networks. *to appear in European Symposium on Algorithms (ESA)*, Sept. 2002. Technical Report No. LA-UR-02-2427, Los Alamos National Laboratory.

[23] C. Barrett, R. Jacob, and M. Marathe. Formal Language Constrained Path Problems. *SIAM J. Computing*, 30(3):809–837, 2001.

[24] R. Jacob, M. Marathe, and K. Nagel. A Computational Study of Routing Algorithms for Realistic Transportation Networks. *ACM J. Experimental Algorithmics*, 4:Article 6, 1999. http://www.jea.acm.org/1999/JacobRouting/.

[25] R. Sedgewick, *Algorithms*, Addison-Wesley (1988).

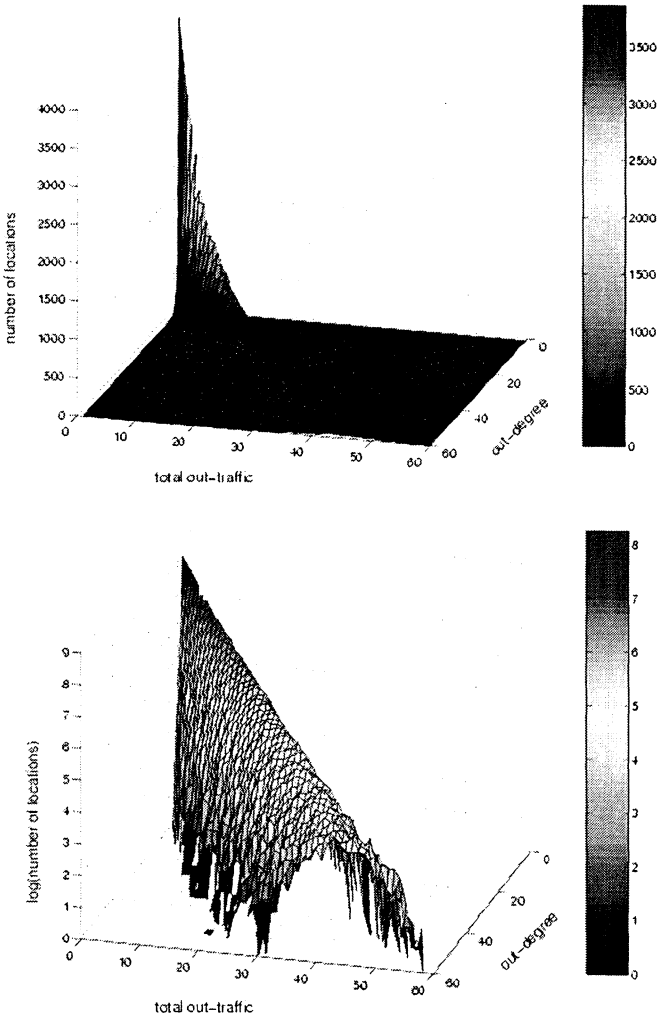

Figure 8: (a)Joint distribution $F(k, v)$ plot (b)log_e density of $F(k, v)$ plot between the out-degree k and the total out-traffic v in the location-based network of the city of Portland.