TOWARDS GENERATIVE ACTIVITY-BASED MODELS FOR LARGE-SCALE SOCIO-TECHNICAL SIMULATIONS (EXTENDED ABSTRACT)

G. ISTRATE, Los Alamos National Laboratory, Los Alamos, NM
A. HANSSON,* Los Alamos National Laboratory, Los Alamos, NM
C. D. TALLMAN, Los Alamos National Laboratory, Los Alamos, NM
L. CUELLAR-HENGARTNER, Los Alamos National Laboratory, Los Alamos, NM
N. HENGARTNER, Los Alamos National Laboratory, Los Alamos, NM

ABSTRACT

An approach to activity modeling based on the theory of random graphs has been proposed in (Eubank et al., 2004a). The approach does not represent temporal duration and activity type information, and does not yield a full activity generator. We present a number of theoretical concepts and experimental results supporting this goal. They include (i) the use of multi-labeled bipartite graphs to represent the missing information, (ii) the automatic inference of the label set via temporal clustering of activities, (iii) we highlight the existence and role of clustering and community structure, a feature of real activity sets that needs to be incorporated in the model.

Keywords: Activity modeling, random graphs, cultural similarity, computational social theory

INTRODUCTION

Activity-based approaches (Timmermans, 2005) emerged in the 1970s, and are becoming increasingly popular in Computational Social Theory, particularly in areas such as travel planning and forecasting, disease modeling and mitigation, and in several other problems related to homeland security. Indeed, since it is often the case (for instance when modeling disease transmission) that the social interaction network is the determining factor for the overall social dynamics, one could in principle replace (and greatly speed-up) agent-based simulations by simulating the underlying dynamics directly on the social network. Analytical models of the activity network have the additional benefit of leading (at least in principle) to *generic methods* for modeling social dynamics.

Developing generic models of human behavior for urban infrastructure simulations has a number of potential benefits to simulation-supported decision-making:

- Such generic models could replace nonparametric, purely data-driven activity-generation in social agent simulations, enabling rapid response, quick turnaround studies.
- Parametric models are more readily transferred to situations when detailed data is incomplete or lacking, and can overcome the inconsistencies of commercial and publicly available data sets.

*Corresponding author address: Anders Hansson, Discrete Simulation Sciences (CCS-5), Los Alamos National Laboratory, P.O. Box 1663, MS M997, Los Alamos, NM 87545; e-mail: hansson@lanl.gov.

- Such models facilitate the computational assessment of the robustness of various general guidelines (normative properties or policies) with respect to variations in the quantitative properties of the simulations.

A recent article in *Nature* (Eubank et al., 2004a) proposes an approach to activity modeling based on ideas from the theory of random networks (Newman et al., 2006). The authors investigate properties of the bipartite people-location networks arising from computational runs of the multiagent simulation EpiSims.[1] The networks arising from EpiSims runs naturally encode physical contact patterns that arise from movements of individuals between specific locations. As shown in (Eubank et al., 2004a), the structural properties of these networks (e.g., the so-called *graph expansion*) have significant implications for the efficiency of disease mitigation and control. Experimental studies have further shown that the Aiello-Chung-Lu (ACL) model from random graph theory captures a number of structural characteristics of this activity-induced physical contact graph.

The basis for the use of random graph models in activity modeling is encapsulated in the following two hypotheses:

- The *low-dimension social modeling hypothesis:* In a nutshell, *macroscopic characteristics of the bipartite graph generated from activity schedules can be described by a small number of parameters*.
- The *cultural similarity hypothesis:* The macroscopic characteristics of urban environments show a significant degree of correlation (at least for large cities in the United States).

The two hypotheses were stated in an unpublished manuscript (Barrett et al., 2004), with a number of measurements to support them. However, the results in (Barrett et al., 2004) do not have significant implications on the definition of synthetic activity generators as few modeling guidelines are provided.

Our long-term research objective is to complete the random-graph based approach to a full-fledged generative model of activities, incorporating temporal and activity-type information (components missing in the preliminary model proposed in (Eubank et al., 2004a)). In this paper, we present information supporting this goal:

- First, we show how to generalize the ACL model to a bipartite random graph model that incorporates multi-label information. The set of labels can then be chosen to encode information about both activity types and their start- and end-times.
- We then show how the set of labels can be inferred from real survey data by clustering.
- Finally, we give evidence for the low-dimensional nature of activity data. In particular, we show how different demographic communities can be inferred from activity data.

[1] EpiSims is a large-scale individual-based epidemiology simulation developed at Los Alamos National Laboratory (Barrett et al., 2005), based on detailed census, land-use, and population-mobility sample data.

RANDOM NETWORK MODELS OF ACTIVITY DATA AND THEIR DRAWBACKS

The results in (Eubank et al., 2004a) are based on modeling activities using *bipartite graphs (networks):*

Definition 1 *A* bipartite network $G = (V, E)$ *is a network such that its vertex set* V *can be partitioned into nonempty subsets* $V = V_1 \cup V_2$ *such that every edge in the edge set* E *of* G *has one endpoint in* V_1 *and one endpoint in* V_2.

Activities can be naturally modeled by bipartite graphs $G = (P, L, E)$ where P is the set of *people*, L is the set of *locations*, and for every person $p \in P$ and any activity a of P performed at location $l \in L$ we add an edge (p, l) to graph G. The edge can be endowed with additional information, such as the start- and end-times of the activity, the activity type, etc.

The results in (Eubank et al., 2004a) draw on existing literature in random graph theory, adapted to bipartite graphs. A critical measure to capture is *degree distribution*. This measure differentiates epidemiology approaches based on random graphs from the more common *SIR models*, which assume very simple structure of the social network, given by uniform or block mixing. Being able to capture arbitrary degree distributions gives rise to the most natural model of this type, the so-called *configuration model*. It is specified by a distribution of degrees $(d_1, \ldots, d_n)$. Sampling from this model is reasonably easy: for every index i we create d_i copies of a node, connected to all nodes arising from a different index. One then considers a random perfect matching in this graph, and contract the copies of each node. With constant probability this gives rise to a simple graph (one without duplicated edges).

Since the generation and analysis of random samples from the configuration model is reasonably complicated, the paper (Eubank et al., 2004a) used a "relaxed" version of this model, the so-called Aiello-Chung-Lu random graph model, more precisely its adaptation to bipartite networks:

Definition 2 *Let* $D_1 = (d_{1,1}, \ldots, d_{1,p})$ *and* $D_2 = (d_{2,1}, \ldots, d_{2,l})$ *be two sequences of integers with equal sum. A sample of the ACL bipartite graph with degree distributions* D_1 *and* D_2 *is a bipartite graph with* $p + l$ *vertices labeled* $\{P_1, \ldots, P_p\}$ *and* $\{L_1, \ldots, L_l\}$. *Any two nodes* P_i *and* L_j *are connected independently at random with probability* $p_{i,j} = \frac{d_{1,i} \cdot d_{2,j}}{\sigma}$, *where* σ *is the proportionality factor* $\sigma = \sum_{i,j} d_{1,i} \cdot d_{2,j}$ *(such that resulting probabilities add up to one).*

In (Eubank et al., 2006) it was shown that a number of structural characteristics of the people-location network in EpiSims can be captured to a reasonable degree of accuracy by a random ACL graph with exponential distribution of degrees on the people side and a power law degree on the locations side.

Nevertheless, a random model such as the one outlined above has a number of drawbacks:

- The model only captures activity properties on a timescale of 24 hours (or whatever time scale is used for recording activities). It does not incorporate information on activity types and their start- and end-times, and thus cannot be used for activity generation.

- It does not capture *community structure*, as displayed in real networks. Indeed, the probability of an edge (in an ACL random graph) depends only on the degree of the two nodes. In contrast, in real urban environments the activities that people undertake are highly clustered, reflecting both *geographic* and *demographic* locality, and this is potentially important for many problems (e.g. disease propagation).

Multi-labeled Networks and the Temporal Disaggregation of Activity Data

In this section we give a generalization of the ACL random graph that is compatible with the task of activity modeling. The basis for the generalization is the following observation: suppose that instead of edges of a single type we "color" each edge with a symbol chosen from a fixed set of *labels*. The coloring is such that no more than one edge adjacent to a person node is colored with a given label. Then we can define, for every person or location node n and every label μ, the *degree of node n with respect to label μ*, as the number of edges in graph G that are adjacent to node n and labeled μ. By the convention we imposed that the degree of any person node with respect to a label μ is either zero or one. In contrast, there is no such restriction on location nodes.

Suppose now that we define the set of labels, and we color the activity graph such that:

- *Labels refine activity types*. That is, for any label l there exists an unique activity type A (e.g., home or work) such that all edges labeled l correspond to activities of type A.
- *Labels record temporal information*. That is, all activities labeled l are "clustered" with respect to activity start- and end-times. These activities correspond to approximately similar time periods.

One can now reduce the problem of activity generation to one of generating random bipartite graphs with similar label degree distributions. Formally we want to generate random samples from the model specified as follows:

Definition 3 *Let $W = \{l_1, \ldots, l_k\}$ be a set of labels with k elements, and let $D_1 = (\overline{d_{1,1}}, \ldots, \overline{d_{1,p}})$ and $D_2 = (\overline{d_{2,1}}, \ldots, \overline{d_{2,l}})$ be two sets of k-tuples of integers with equal sum.*

A multi-labeled bipartite graph with generalized degree distributions D_1, D_2 *is a bipartite graph $V = (V_1, V_2, E)$ with edges labeled with labels in W such that for all $j \in 1 \ldots k$*

- *For all $i = 1, \ldots, p$, the number of edges adjacent to the ith element of V_1 that are colored with label l_j is $d_{1,j}$.*
- *For all $i = 1, \ldots, l$, the number of edges adjacent to the ith element of V_2 that are colored with label l_j is $d_{2,j}$.*

One can easily see the model in the previous definition as a generalization of the configuration model: to each vertex add a number of colored stubs, with their number specified by the proper degree. Create then, separately for each color, a random matching between stubs on the left and right-hand side of the graph.

The kind of labels we are going to use will aim to reflect the additional information available in the activity data that is not captured by the degree information. For instance a label could

correspond to activity type "work without interruption, from approximately 9 to 5." In general, however, we will have to use more complicated encodings in cases where several activities are highly correlated. For instance, "work" interrupted by "lunch" activity will show in our data as two work activities with different time intervals taking place (in most cases) at the same location. It is, therefore, useful to introduce a single label that will represent both work before and after lunch.

The set of activities of a single person respects temporal disjointness: at most one activity can take place at any one time. This constraint will be reflected by a corresponding constraint on the labels.

Of course the generalization of the configuration model described above still suffers from a number of drawbacks:

- It is impractical to completely specify the set of label degrees. Rather, we would like to model the degree distributions.
- The model still does not incorporate community structure.
- It is not clear that a set of labels can be inferred from the data.

In the next sections we address these issues. First, we show that labels can often be inferred from temporal clustering of activities. We present preliminary results on identifying clusters of activity patterns.

STRUCTURAL ANALYSIS OF ACTIVITY DATA IN REAL URBAN ENVIRONMENTS

We now present the results of a preliminary investigation of activity data arising from three surveys: a national-level household survey, as well as two household survey for the urban areas of Chicago and Houston.

The National Household Transportation Survey (NHTS) is a nationwide survey of travel patterns taken in 2001 (U.S. Department of Transportation, 2001). The table DAYPUB, which records trips for each person on a particular day, was converted from trips to activities by looking at the intervals between trips. All activities were selected except for those persons with flags indicating that they were out of town, and persons who had missing time information for their activities.

There were $761{,}811$ activity records in total. In all cases we eliminated any activity that crossed the maximum time, which was $1{,}740$ minutes (29 hours) from the start of the survey. We binned the start and end time activity pairs into a table of 30-minute intervals, truncating intermediate times to the previous 30 minutes. We normalized the data, dividing by the sum of the entries in the table to create a probability density graph, and plotted it with 24 color breaks ranging from 2×10^{-15} to 2×10^{-3}. We added contour lines at 1.25×10^{-3}, 2.5×10^{-3}, and 5×10^{-3}, labeling them as 1, 2, and 4 respectively. These levels are arbitrary; they help to show the shape of the probability density surface.

The Chicago data is from the 1990 Chicago Area Transportation Study (CATS) (Ghislandi, 1990). The activity records were taken from the Trip files table, converted to activities as in the

NHTS table above. There were 189,253 activity records in total. We binned, normalized, and plotted the data as in the NHTS case above. The primary difference is that for this survey the time was counted from 4:00 a.m. We added back four hours in the plot to show comparable results.

The Houston data is from a 1994 Household Travel Survey by the Houston / Galveston Area Council (Houston-Galveston Area Council, 2001). The activity records were taken from the activity table ACTIVITY.ORG. There were 29,045 activity records in total. The 541 work-related activities, the 338 other activities, and the 196 college activities were rejected for being too small a sample size. We binned, normalized, and plotted the data as in the NHTS case above.

The activity types employed in the three surveys do not completely overlap. Table 1 presents the breakout of activities per activity type for each of the three survey. An "NR" (Not Recorded) entry signifies that an activity type with that name was not recorded in the study.

We will now present probability density graphs for start- and end-times of four different activities: home, work, education, and serve passenger. As we will see, the graphs will all support the notion of a cultural similarity. College activities were not explicitly recorded in the CATS data; they were instead added to school activities. In order to compare the three sets of survey data, we therefore merged the school and college activities of the NHTS and Houston surveys, respectively. Once aggregated, these two activity types are classified as "education."

TABLE 1: Activity counts in sample data

Activity type	NTHS	Chicago	Houston
Work	56,843	28,235	3,308
Work-related trips	14,508	6,781	541
Home	220,413	90,168	8,321
School	14,739	4,526	1,418
College	2,757	NR	196
Retail	80,348	16,341	2,298
Serve passenger	68,398	10,759	2,200
Visit	26,893	NR	NR
Services	NR	NR	1,326
Medical	9,021	NR	NR
Recreation	41,496	7,700	2,911
Banking	16,443	2,983	NR
Meal	36,410	15,200	NR
Daycare	2,926	NR	NR
Other	2,926	13,520	338
TOTAL	761,811	189,253	29,045

Clustering of Home Activity Data

Figure 6 shows the probability density graphs of the home activity start- and end-times for the NHTS, Chicago, and Houston surveys. The three graphs display the following common features:

- There is a distinct cluster at the very left of each graph, which simply reflects that people are at home at the beginning of the survey (i.e., in the early morning). It is also seen how most people leave their home between 6–9 a.m.
- There are two more clusters, corresponding to lunch, which takes place around noon, and evening activities, which peaks some time between 5–6 p.m. As expected, the lunch cluster is closer to the diagonal, as this indicates a shorter activity (generally less than an hour).

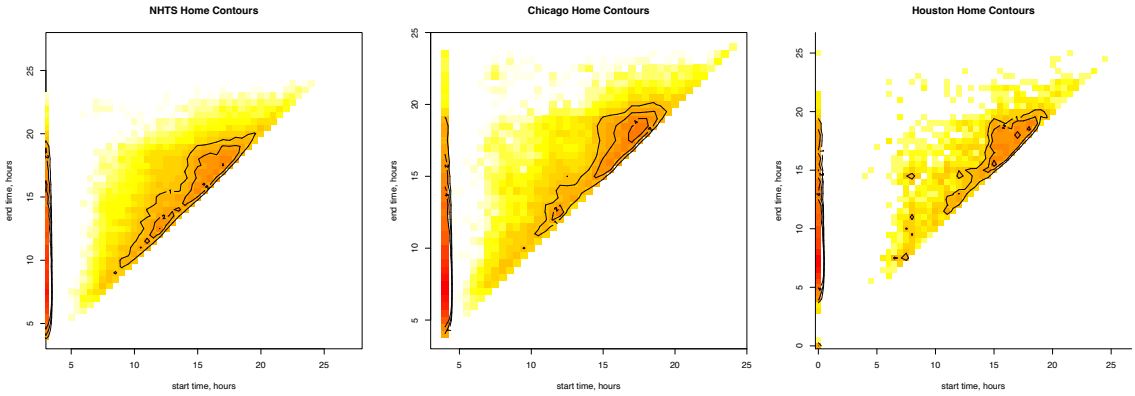


FIGURE 1: Probability densities of the home activity start- and end-times for (a) NHTS, (b) Chicago, and (c) Houston

Clustering of Work Activity Data

Figure 2 shows the probability density graphs of the work activity start- and end-times for the NHTS, Chicago, and Houston surveys. The three graphs display the following common features:

- The most dominant cluster is farthest away from the diagonal, and this corresponds to an uninterrupted workday, starting between 6–9 a.m. People in this cluster were generally recording a single work activity throughout the day.
- By instead considering people who (in general) recorded two work activities throughout the day, we can locate the two clusters closer to the diagonal. Like the first cluster, these clusters are also very distinct, and they correspond to work before and after lunch, respectively. It can be seen how the start-time of the first cluster (before lunch work) and the end-time of the second cluster (after lunch work) naturally peak at the same time as the start- and end-times of the uninterrupted work activity. The first work activity ends around noon and the second starts around 1 p.m.
- One can also see some smaller clusters that correspond to shift workers who start their day in the afternoon and work until the evening. (Some shift work activities for Houston are apparently cut off as they appear at the very top of the graph.)

Let us provide two more remarks concerning the work activity data. The Houston survey is smaller, and this fact explains why the Houston graph shows less dispersion than the other two graphs. Finally, we conclude that work makes a good candidate for an *anchor activity* (for adults), i.e., once the work activity pattern of a given individual has been established, one can readily sample the start- and end-times of his/her remaining activities (e.g., a person who interrupts his/her work for a lunch break could potentially go home for lunch).

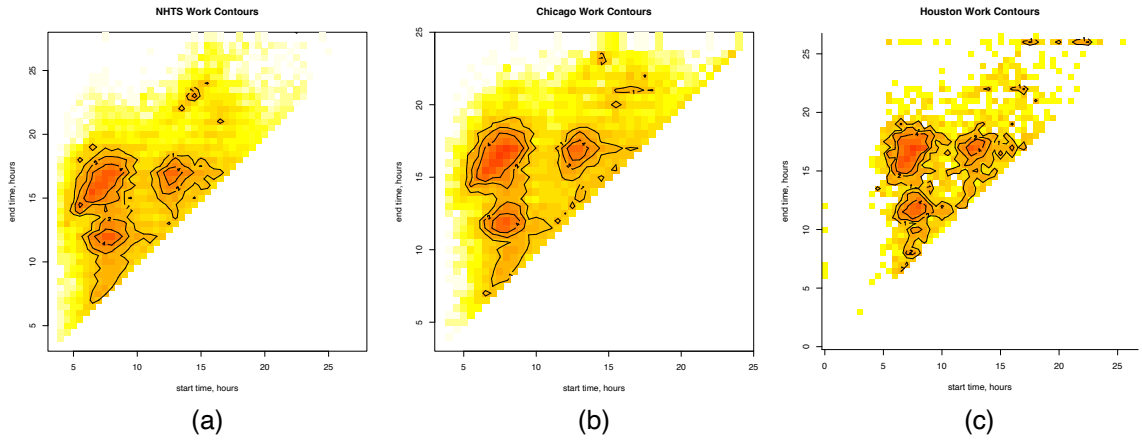


FIGURE 2: Probability densities of the work activity start- and end-times for (a) NHTS, (b) Chicago, and (c) Houston

Clustering of Education Activity Data

Figure 3 shows the probability density graphs of the education activity start- and end-times for the NHTS, Chicago, and Houston surveys. The three graphs display the following common features:

- The education activities appear to be somewhat more fragmented than work activities (for example, see the Chicago data). We do not attribute this behavior to a smaller sample size, but instead hypothesize that students engage in various other activities between classes.
- The three clusters that was seen for work activities can also be identified here—although not as easily: uninterrupted education that starts in the morning and ends in the afternoon, as well as a education before and after lunch (see the NHTS graph).
- One can also see a cluster corresponding to evening classes. This cluster is particularly evident in the Chicago graph, which perhaps could be explained by the rich educational choice of the Chicago metropolitan area (in contrast to the NHTS data in which this effect is somewhat neutralized by rural areas).
- Finally, we observe that there are virtually no education activities before 7 a.m. in the morning, as expected.

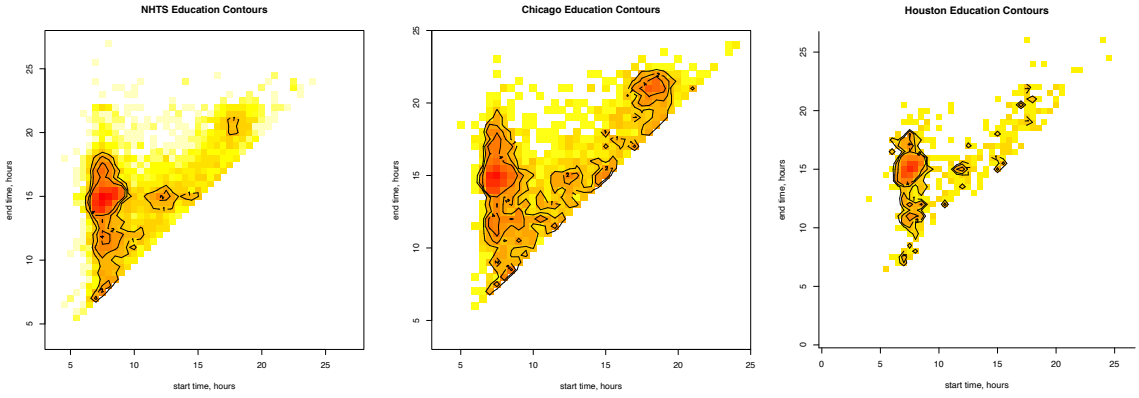


FIGURE 3: Probability densities of the education activity start- and end-times for (a) NHTS, (b) Chicago, and (c) Houston

Clustering of Serve Passenger Activity Data

Figure 4 shows the probability density graphs of the serve passenger activity start- and end-times for the NHTS, Chicago, and Houston surveys. The three graphs display the following common features:

- The serve passenger activity type is more or less evenly distributed throughout the day from around 7 a.m. until around 9 p.m. Moreover, the activities tend to be short (the cluster is concentrated close to the diagonal of the graph).
- One could identify two short activity dips (most apparent in the Chicago data): the first around 10 a.m. and the second around 1 p.m. At these points in time, most people have generally arrived at their work, either in the morning, or after lunch, and the demand for the serve passenger activity naturally drops for a while.

CLUSTERING AND IDENTIFICATION OF ACTIVITY PATTERNS FROM DATA

The previous section has provided evidence for the cultural similarity of two urban areas with respect to one national survey. In this section we provide evidence for the existence of natural community structure. In this extended abstract we do not address the full problem of community detection based on concepts such as modularity. Rather, our goal is to cluster activity sets based on common patterns.

The data-set we used for this purpose was a subset of the Chicago data-set, as used in the

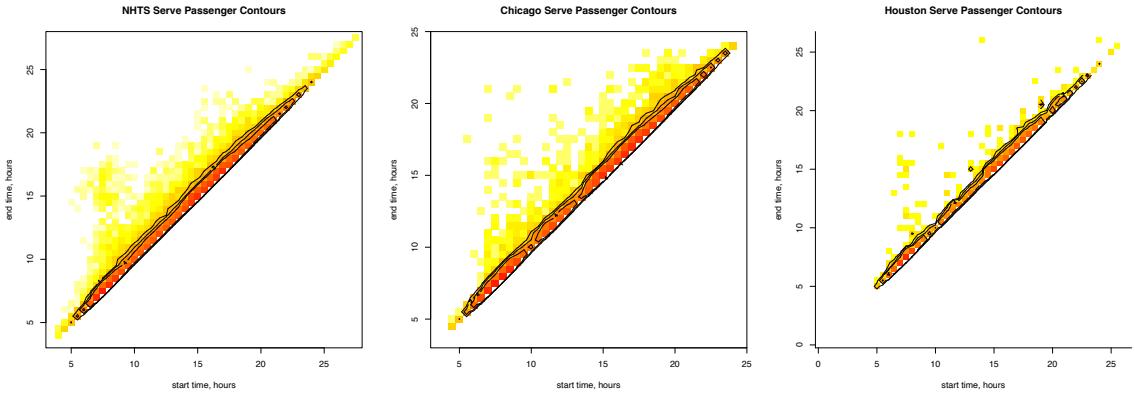


FIGURE 4: Probability densities of the serve passenger activity start- and end-times for (a) NHTS, (b) Chicago, and (c) Houston

simulations Transims and Episims. A sample of 0.1% of the individuals in the simulation was used, resulting in $4{,}365$ individuals. Table 2 displays the percentage of people performing the given activity type at least once.

TABLE 2: Percentage of people performing a specific activity

Home	Work	School	Retail	Other	Serve p.	College
99.8	45.8	20.3	27.7	50.5	11.9	4.1

Next we performed a Principal Component Analysis (PCA) on the sets of individual activities. We create a matrix, whose rows are people and whose columns correspond to activity types. The entries of the matrix are zero, except for the cases when the given person performs the corresponding activity, in which case it is equal to one. PCA tries to find a space generated by the PC's (vectors) such that the distance of the observations (in this case, the $4{,}365$ vectors of 0s and 1s) to the space generated by the PC's is minimized. It does this by iteratively seeks a subspace such that the projection onto that subspace has maximal variance. The method can often reveal clusters in the data. The result of the PCA, presented in Table 3, show that clearly the first three PC's are enough to explain a significant amount (76–84%) of the variance in the input data. Table 4 presents the coefficients in the linear combination of the $0/1$ activity variables that specify the PC's, e.g. the equation of PC_1 is:

$$\begin{aligned} -0.759 \cdot (\text{Home}) - 0.378 \cdot (\text{Work}) - 0.144 \cdot (\text{School}) - 0.235 \cdot (\text{Retail}) \\ -0.44 \cdot (\text{Other}) - 0.103 \cdot (\text{Serve passenger}) - 0.027 \cdot (\text{College}). \end{aligned}$$

The important features in this equation and similar ones are the magnitude of the coefficients and quantities with a negative sign. In PC_1 the most important variables are "Home," "Work," and "Other." This suggests that most of the variance in the activity data is given by the

presence or absence of one of these three activity types. For PC_2 the important variables are "Work," "Other," and "School." Also, observe that "Work" has opposite sign to "School" and "Other." So the variance along PC_2 is capturing the group of people that work, but neither go to school nor do "other" activities, or those that are in the opposite situation. Similarly, the variance along PC_3 captures people going to school that do not do retail activities and vice-versa, people that do retail but do not go to school. The results of the PCA are also displayed in Figure 5, where the projection of activity points is presented with respect to each pair of PC's.

We have also performed PCA on the *total time* spent doing the activities. That is, for each person we keep track of how much time he/she spent doing each of the activities. This amount becomes the input to a matrix indexed by people (for the rows) and activity types (for the columns). The relative importance of the PC's in explaining variance in the data is presented in Table 5. It is easy to see that most variability in the data is explained by the first four PC's.

TABLE 3: Importance of PCA components for activity data

Importance of components:	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Standard deviation	1.288	0.578	0.462	0.4319	0.3193	0.2612	0.1935
Proportion of variance	0.638	0.128	0.082	0.0717	0.0392	0.0262	0.0144
Cumulative proportion	0.638	0.766	0.848	0.9202	0.9594	0.9856	1.0000

TABLE 4: PCA coefficients for activity data

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Home	−0.759	0.035	−0.230	0.220	−0.056	−0.540	0.164
Work	−0.378	0.774	0.060	−0.165	−0.042	0.466	−0.091
School	−0.144	−0.298	−0.677	0.319	0.037	0.553	−0.150
Retail	−0.235	−0.212	0.664	0.611	−0.082	0.273	−0.064
Other	−0.440	−0.515	0.192	−0.670	−0.068	0.219	−0.058
Serve passenger	−0.103	−0.008	0.083	−0.002	0.991	0.009	0.010
College	−0.030	0.024	0.006	0.003	0.009	−0.252	−0.967

The coefficients of the PCA are listed in Table 6. In this case the most important variables for PC_1 are "Home" and "Work," and then "Other" and "School" (but much less than in the PCA for activities). The second PC is most influenced by the "Work," "Other" and "School" variables. Note the opposite signs: people working spend less time doing "other" activities and/or at school. PC_3, on the other hand, depicts people that do "other" activities and work.

The percentage of time people spent in each activity is presented in Table 7. Not surprisingly, the time spent at home and at work are most important. The "other" activities category, on the other hand, does not seem to be so important in the PCA.

In Figure 6 (b) we analyze more closely the relation between the first three PC's. Looking at the PC_1 vs. the PC_2 plot unveils two clusters of points along the wedges of the triangle. They

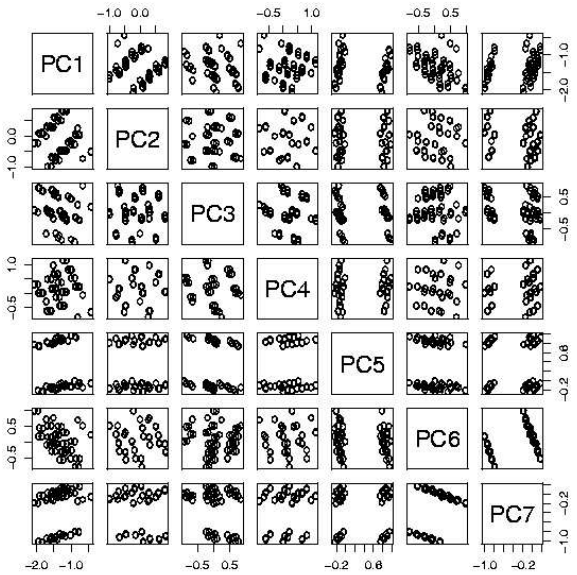


FIGURE 5: PCA for activity data

TABLE 5: Relative contribution of PC's to explaining variance in duration data

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Standard deviation	19.70	4.98	4.12	3.36	1.45	0.91771	0.66118
Proportion of variance	0.87	0.06	0.04	0.03	0.01	0.00189	0.00098
Cumulative proportion	0.87	0.93	0.97	0.99	0.997	0.99902	1.00000

mostly correspond to a distinction into two distinct group of people according to the time spent working. Indeed, for PC_2, work is the only activity with a negative coefficient. People with a strong negative value of PC_2 are those spending a significant amount of time in work activities. In contrast, people with a positive value in the PC_2 component spend more time in the school and "other" activities.

We next present results on the independence and correlation of activities. Table 8 gives the total number of people that perform both of a given pair of activities. The diagonal counts people that perform the given activity at least once. Table 9 presents the corresponding joint probabilities of the main activity categories. Observe that activity home is independent of all other activities (see also the correlation below). For example, the probability of working and doing retailing is 0.17. The probability of working is 0.46.

We now compute the conditional probabilities, given by $p(i|j) = p(i,j)/p(j,j)$. The results are presented in Table 10. So for example the probability of working, given that the person goes retailing is 0.999. Or probability of retailing, giving that one works, is 0.2. A further measure

TABLE 6: PCA coefficients for duration data

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Home	-0.981	0.091	0.148	0.080	0.023	0.010	0.003
Work	-0.165	-0.853	-0.431	-0.241	-0.024	-0.008	0.000
School	-0.062	0.332	-0.109	-0.935	-0.007	-0.015	-0.002
Retail	-0.018	0.022	0.012	0.016	-0.999	-0.017	0.001
Other	-0.076	0.391	-0.883	0.247	0.003	-0.005	0.000
Serve passenger	-0.003	0.000	0.001	0.002	-0.001	0.014	-1.000
College	-0.007	0.000	0.011	0.015	0.017	-1.000	-0.014

TABLE 7: Percentage of time spent in various activities

Home	Work	School	Retail	Other	Serve p.	College
66.4	13.0	5.6	1.5	7.4	0.3	0.5

of activity independence is given by the amount $p(i|j) - p(i,i)$. A value close to zero in Table 11 signifies independence.

A different indicator of independence, the statistical correlation between activities, is presented in Table 12. Home activity is independent of every other activity, and college seems to be the activity less correlated with all the others. Also observe that the correlations between activities is low, somewhere between -0.4 and 0.1.

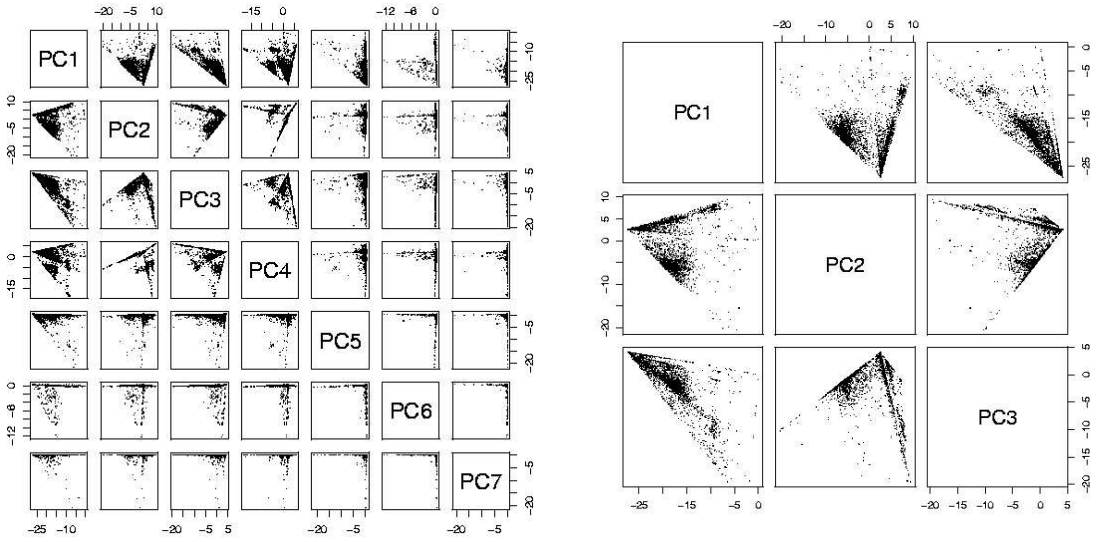
The conclusion of the preliminary investigations of this section suggest that clustering activity patterns is likely to be significant in a full-fledged model. Ignoring such clustering destroys all correlations between activities.

CONCLUSIONS

We have outlined an approach that can lead in principle to a synthetic model of activities in real data that extends the approach in (Eubank et al., 2004a). Significant work remains to be done. For instance, the clustering analysis in the previous section is most natural on the set of "labels," the refinements of activity types we highlighted. Such an analysis (combined with community inference methods) would hopefully result in a hierarchical random model that can represent locality. The existence of fast generation algorithms for our network models (similar to the work in (Eubank et al., 2004b)) is an interesting algorithmic problem. Finally, the accuracy of random models of activity data should be verified more thouroughly.

ACKNOWLEDGMENTS

This work has been supported by the Department of Energy under contract W-705-ENG-36 and by the Department of Homeland Security under the National Infrastructure Simulation and Analysis Center (NISAC) program.



(a) (b)

FIGURE 6: (a) Clustering by pairs of PCA components, (b) The first three PCA components

TABLE 8: Joint activity counts in sample data

	Home	Work	School	Retail	Other	Serve p.	College
Home	4356	2000	886	1206	2195	520	179
Work	0	2001	56	400	757	260	88
School	0	0	886	123	424	75	0
Retail	0	0	0	1207	715	192	38
Other	0	0	0	0	2204	319	69
Serve passenger	0	0	0	0	0	520	24
College	0	0	0	0	0	0	179

REFERENCES

Barrett et al. (2004). A tale of two cities. (unpublished manuscript).

Barrett, C. L., Eubank, S. G., and Smith, J. P. (2005). If smallpox strikes Portland ... *Scientific American*.

Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004a). Modeling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184.

Eubank, S., Kumar, V. S. A., Marathe, M., Srinivasan, A., and Wang, N. (2004b). Structural and algorithmic aspects of massive social networks. In *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 711–720.

Eubank, S., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., and Wang, N. (2006). Structure

TABLE 9: Joint probabilities of main activity types

	Home	Work	School	Retail	Other	Serve p.	College
Home	1.000	0.459	0.203	0.277	0.504	0.119	0.041
Work	0.459	0.459	0.013	0.092	0.174	0.060	0.020
School	0.203	0.013	0.203	0.028	0.097	0.017	0.000
Retail	0.277	0.092	0.028	0.277	0.164	0.044	0.009
Other	0.504	0.174	0.097	0.164	0.506	0.073	0.016
Serve passenger	0.119	0.060	0.017	0.044	0.073	0.119	0.006
College	0.041	0.020	0.000	0.009	0.016	0.006	0.041

TABLE 10: Conditional probabilities of main activity types

	Home	Work	School	Retail	Other	Serve p.	College
Home	1.000	1.000	1.000	0.999	0.996	1.000	1.000
Work	0.459	1.000	0.063	0.331	0.343	0.500	0.492
School	0.203	0.028	1.000	0.102	0.192	0.144	0.000
Retail	0.277	0.200	0.139	1.000	0.324	0.369	0.212
Other	0.504	0.378	0.479	0.592	1.000	0.613	0.385
Serve passenger	0.119	0.130	0.085	0.159	0.145	1.000	0.134
College	0.041	0.044	0.000	0.031	0.031	0.046	1.000

of social contact networks and their impact on epidemics. In Abello, J. and Cormode, G., editors, *Discrete Models in Epidemiology*, volume 70 of *DIMACS-AMS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society.

Ghislandi, A. C. (1990). CATS 1990 household travel survey, technical documentation for the household, person, and trip files. Technical report, Chicago Area Transportation Study Information Services Division, Chicago, IL.

Newman, M., Barabasi, A. L., and Watts, D., editors (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

Houston-Galveston Area Council (2001). Regional travel models. 1995 model validation and documentation report. Technical report, http://www.h-gac.com/HGAC/home/Default.htm.

U.S. Department of Transportation (2001). The national household travel survey (NHTS). Technical report, Center for Transportation Analysis, Oak Ridge National Laboratory, http://nhts.ornl.gov/2001/index.shtml.

Timmermans, H., editor (2005). *Progress in Activity-Based Analysis*. Elsevier.

TABLE 11: Degree of dependence between main activity types

	Home	Work	School	Retail	Other	Serve p.	College
Home	0.000	0.000	0.000	−0.001	−0.004	0.000	0.000
Work	0.000	0.541	−0.396	−0.128	−0.116	0.041	0.032
School	0.000	−0.175	0.797	−0.101	−0.011	−0.059	−0.203
Retail	0.000	−0.077	−0.138	0.723	0.047	0.092	−0.065
Other	−0.002	−0.128	−0.027	0.086	0.494	0.107	−0.120
Serve passenger	0.000	0.011	−0.035	0.040	0.025	0.881	0.015
College	0.000	0.003	−0.041	−0.010	−0.010	0.005	0.959

TABLE 12: Statistical correlation of main activity types

	Home	Work	School	Retail	Other	Serve p.	College
Home	0	0.000	0.000	0.000	0.000	0.000	0.000
Work	0	1.000	−0.402	−0.159	−0.235	0.030	0.013
School	0	−0.402	1.000	-0.156	-0.028	-0.054	−0.105
Retail	0	−0.159	−0.156	1.000	0.107	0.076	−0.030
Other	0	−0.235	−0.028	0.107	1.000	0.079	−0.050
Serve passenger	0	0.030	−0.054	0.076	0.079	1.000	0.009
College	0	0.013	−0.105	−0.030	−0.050	0.009	1.000