Sampling Stable Properties of Massive Track Datasets

Christopher I. Connolly
Artificial Intelligence Center
SRI International
Menlo Park, CA USA
+01 650 859-5022
connolly@ai.sri.com

J. Brian Burns
Artificial Intelligence Center
SRI International
Menlo Park, CA USA
+01 650 859-5326
burns@ai.sri.com

Hung H. Bui
Artificial Intelligence Center
SRI International
Menlo Park, CA USA
+01 650 859-5352
bui@ai.sri.com

ABSTRACT
Analysis of massive track datasets is a challenging problem, especially when examining n-way relations inherent in social networks. In this paper, we explore ways in which stable properties of sensor observations can be extracted and visualized using a statistical sampling of features from a very large track dataset, using very little ground truth or outside knowledge. Special attention is given to methods that are likely to scale well beyond the size of the Mitsubishi dataset.

Categories and Subject Descriptors
I.5.4 [**Pattern Recognition**]: Clustering – *algorithms, similarity measures.* The ACM Computing Classification Scheme: http://www.acm.org/class/1998/

General Terms
Algorithms, Measurement, Performance, Experimentation.

Keywords
Sampling, clustering, tracking.

1. INTRODUCTION
We wish to explore ways in which stable (at least in the piecewise sense) properties of sensed activity can be extracted from massive datasets. Social interactions in an office setting represent one class of activity that, in most cases, can be stable over long periods of time (months or years). Social networks are relatively long-term relationships that may arise through common interests, including work projects, hobbies, or simply common destinations in a physical environment. For very large data sets, it may be unrealistic to rely on exhaustive analysis to uncover these relationships. We therefore focus on statistical methods requiring only modest computing resources to find groupings of interest in the data.

Social network analysis is an area that is receiving increased interest. As sensors proliferate, an increasing volume of data is generated that, implicitly at least, contains information about social groupings among sensed individuals [1]. Knowing these groups can be important for a variety of reasons. For example, the DARPA CALO project [2] exploits knowledge of social groups to anticipate user needs and scheduling conflicts. Physical security and safety applications can use knowledge of social groups to identify unusual, dangerous or threatening behavior.

Workshop on Massive Datasets '07, November 1–2, 2007, Nagoya, Japan.

The MERL (Mitsubishi Electric Research Lab) dataset [3] is an excellent example of a massive dataset that can be used to expose social relations. This is a database of activations of motion sensors distributed through the MERL environment, recorded during the course of 1 year. A high volume of data is generated each day, resulting in a dataset that is difficult to process exhaustively in a short amount of time. The situation is further complicated when one wishes to examine n-way social networks inherent in such a dataset. The data structures of interest, as defined in the MERL dataset, are *tracklets*: temporally ordered chains of adjacent sensor activations, and *tracklet graphs*: partially ordered graphs of tracklets that describe the possible trajectories of one or more movers in the environment.

1.1 Importance Sampling and Accumulation
Social networks and personal attributes are assumed to be stationary properties of the dataset, or at least stationary within some time interval in the dataset. Social attributes can only be discovered when the individuals involved are active. In addition, there is a potential overhead incurred in drawing samples from a possibly remote database. These two factors suggest the use of temporal importance sampling that is biased toward times of high activity [4].

A bootstrapping approach can be used to generate the desired distribution. A timeline sampler with a uniform distribution is used first to draw time samples. A timeline accumulator is constructed whose histogram maps onto a duration of one week, with a bin size of one minute. Using the timeline sampler, sensor hits are accumulated modulo one week, so that the accumulator learns a model of activity for the typical week at MERL. Figure 1 shows one example of a uniform sampling run that accumulated statistics over 1.5 million sensor hits. This figure clearly shows the diurnal progression of activity during business days (0 is midnight Sunday), with the expected reduction in activity during nights and weekends. From the distribution shown in Figure 1, we can bootstrap an importance sampler that is biased in favor of times of high activity.

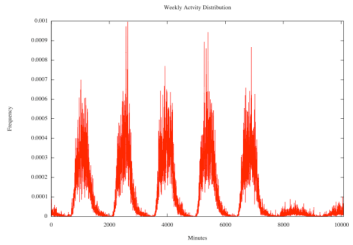


Figure 1: Frequency of sensor hits, accumulated weekly.

1.2 Spatial Sampling and Social Relations

Social relations in groups are characterized in part by the pairwise interactions of members of each group. We expect people who work together to meet and interact, either by visiting offices or by attending meetings. Another potentially observable behavior is walking together (while going to lunch, for example). This might reveal social interactions that are not necessarily work-related, but can be correlated to groups with a common purpose.

We define a *trip* as a tuple (S,**T**) consisting of one source sensor S and all possible terminal sensors $x \in \mathbf{T}$ that are connected to the source through a tracklet graph. The trip represents the possible destinations for a single source sensor in a tracklet graph. During accumulation, each trip adds 1/|**T**| to each of the destination bins.

The visit matrix **A** represents the number of visits made directionally between pairs of sensors, weighted by the number of alternatives found in each case. After sampling converges, each element A_{ij} is a measure of the frequency of travel from sensor i to sensor j. The visit matrix is one simple measure that could reveal information about stable social interactions within the group.

1.3 Entropy

Entropy estimates can be used to assess convergence of samplers and to gauge the relative organization of computed distributions. The most straightforward entropy estimator is the naïve Shannon entropy defined as:

$$H = -\sum_{i=1}^{m} p_i \log p_i$$

where p_i is the frequency corresponding to bin i in our accumulator.

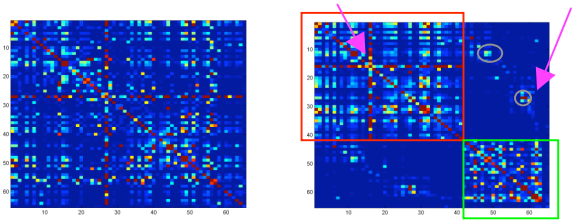
Entropy computation is prone to sampling error, and several estimators have been proposed to address this problem [5,6,7]. On the other hand, entropy estimates generally improve with increased sample size. This suggests that entropy-based analysis of datasets will scale well with increasing database size. Our initial experiments confirmed this in that convergence was fairly rapid, and generally only required a few thousand sampling runs to get stable, low-error values. If anything, the MERL dataset is somewhat smaller than the ideal for entropy-based analysis, since there is a risk of redundant sampling.

Our analysis assumes a stationary dataset. Events can occur that significantly change the interaction patterns in the dataset. For example, an office renovation might require movement of individuals to new offices. If we assume that such events are relatively rare, it may be possible to perform piecewise entropy estimation over separate intervals in the dataset, to segment periods between significant changes in interaction patterns. This is an area of study we intend to pursue further in our own tracking system.

1.4 Segmentation of Groups

We hypothesize that constructs like the office visit matrix contain information about the level of social contact between the office inhabitants, and thus can provide interesting information about social networking, group organization etc. If this is true, groups can be found by segmenting the set of offices into multiple clusters such that visits among each cluster appear much more often than visits across the clusters.

Formally, this can be posed as graph partitioning problem, and a family of methods based on "graph cuts" can be used to yield the clusters. For example, the normalized-cut algorithm [8] directly attempts to find clusters so that visits within clusters are maximized, while visits across clusters are minimized. An example clustering is shown in Figure 2.



(a) (b)

Figure 2: (a) Office visit matrix, (b) Office visit matrix after clustering into 2 groups.

2. Conclusions

We present tools for extracting information about social networks and individuals in the MERL dataset. A more detailed analysis of the MERL dataset is available in [10].

Most importantly, our analysis looks for invariant, global properties of the group under surveillance. Two-way co-occurrences, spectral analysis, and entropy measures can all contribute to understanding the nature of social interactions in the groups. One challenge to our approach is that the association of sensors with individuals is unknown, as is the reliability of the sensors. Both of these factors have a strong impact on any conclusions we can draw about apparently stable interactions.

To improve tractability, especially for massive collections, our approach to analysis relies on importance sampling, for managing dataset size and avoiding excecssive computation. In our experiments, samplers typically ran for only a short time before convergence, suggesting that this approach scales to datasets much larger than the current MERL collection. In fact, entropy-based analysis will generally improve as dataset size increases.

It is also important to acknowledge other approaches to mining data within large datasets. One drawback to sampling is that it can miss possibly important specific events. An alternative approach to analysis of massive datasets is to use context to focus attention on highly constrained subsets of the dataset. When a dataset has sufficient semantic context associated with it

(identities of specific movers, for example), it should be possible to construct highly constrained queries that will reduce the volume of data that needs to be analyzed.

3. REFERENCES

[1] Choudhury, T. and Pentland, A. 2003. Sensing and Modeling Human Networks using the Sociometer, Proc. 7th IEEE Symposium on Wearable Computers.

[2] Mitchell, T., Wang, S., Huang, Y., and Cheyer, A. 2006. Extracting Knowledge about Users' Activities from Raw Workstation Contents. Proc. 21st National Conference on Artificial Intelligence (AAAI '06), July 2006.

[3] Wren, C. R., Ivanov, Y. A., Leigh, D. and Westhues, J. 2007. The MERL motion detector dataset: 2007 Workshop on Massive Datasets. Mitsubishi Electric Research Laboratories Technical Report TR2007-069.

[4] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 2007. *Numerical Recipes: The Art of Scientific Programming*, Cambridge University Press.

[5] Connolly, C. I., and Quam, L. 2007. FREEDIUS: An Open Source Lisp-Based Image Understanding Environment. In Proceedings of the 2007 International Lisp Conference.

[6] Miller, G. and Madow, W. 1954. On the maximum likelihood estimate of the Shannon-Wiener measure of information, Air Force Cambridge Research Center Technical Report 75 (1954) 54-75.

[7] Paninski, L. 2003. Estimation of entropy and mutual information, Neural Computation 15:1191-1253.

[8] Kennel M. and Mees, A. 2002. Context tree modeling of observed symbolic dynamics, Physical Review E, 66:056209.

[9] Shi, J. and Malik, J. 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol 22(8).

[10] Connolly, C. I., Burns, J. B., and Bui, H. H. 2007. Recovering Social Networks From Massive Track Datasets, SRI AI Center Technical Note 564.

Columns on Last Page Should Be Made As Close As Possible to Equal Length