Structured Multinomial Models of Traffic Flow

Juan K. Lin
Department of Statistics
Rutgers University
Piscataway, NJ 08854
jklin@stat.rutgers.edu

Abstract

We investigate various latent variable models of traffic flow. More specifically, we present various structured multinomial mixture models for analyzing source-destination traffic. A "highway" model is formulated which posits highway entrance and exit hubs, and highway traffic between the entrances and exits. The model's structure is based on an analogy with car traffic from a local origin in one city to a destination in another city. The model's parameters correspond to an onramp traffic distribution from sources to highway entrances, the highway traffic distribution between entrances and exits, and an offramp traffic distribution from highway exits to final destinations. The highway traffic model extracts community structure based on source-destination traffic information, but in addition captures the aggregate "highway" traffic between the communities. This important distinction extends the highway traffic analysis beyond clustering and allows it to extract out underlying backbone traffic structure from traffic data. For comparison, we also describe a "hub" traffic model with no highways which has a latent variable structure that has been well studied in the past.

1. INTRODUCTION

Traffic engineering and network design have been extensively studied in the engineering communities. The investigations cover how best to route traffic based on an existing connectivity graph, and optimizing connectivity paths to best fit the traffic. Source-destination traffic matrix estimation has been addressed from a statistical perspective in e.g.[1][2]. Here we present a probabilistic "highway" traffic model of source-destination traffic. Our goal in the analysis is to model both the underlying community structure and the aggregate traffic between communities. Viewing the source-destination traffic matrix as a weighted graph, we seek to discover both tightly connected regions in the graph, and an underlying "highway" backbone structure in the graph.

Our analysis extends latent variable models which have appeared under the names Latent Class Models [3], Aggregate Markov Models [4]-[6], Non-negative Matrix Factorization (NMF)[7], and probabilistic LSA (pLSA)[8]. Many of the recent applications of these models have been in the fields of natural language processing and information retrieval. These latent variable models when applied to source-destination traffic data translate into a "hub" traffic model with only onramp and offramp traffic to latent hubs. The highway traffic latent variable model contains both highway entrance and exit hubs, and highway traffic between them. This allows the model to find both tightly interconnected communities, and the traffic flow between them. In addition to the analysis of source-destination traffic data, the highway traffic model is applicable to the analysis of random walk traffic on a source-destination connectivity graph. In related work, spectral clustering based on finding communities which minimize transitions between different communities has received considerable attention in image segmentation[9][10].

An outline of this paper is as follows. First we describe the highway traffic model and it's relation to a hub traffic model. Section 3 presents comparative analysis of the highway and hub traffic models for the analysis of traffic on an autonomous system connectivity graph and computer skills graph. Section 4 presents a symmetric hub traffic model. The paper concludes with a discussion of some properties of the highway traffic model.

2. Highway and Hub Traffic Models

Consider traffic flow data consisting of n_{ij} counts of traffic from source $X = i$ to destination $X' = j$. We assume that all sources are destinations, and destinations sources. Discrete latent variables H and H' are introduced which characterize the underlying entrance hubs and exit hubs on the highway. We assume that all entrances are exits, and vice versa. Our model of traffic flow consists of onramp traffic from sources to highway entrances, highway traffic from entrances to exits, and offramp traffic from highway exits to destinations. The model assigns a probability of going from source i to destination j of:

$$p(i,j) = \sum_{k,l} \alpha_{ik}\beta_{kl}\gamma_{jl},$$

where $\alpha_{ik} = P(X = i|H = k)$, $\beta_{kl} = P(H = k, H' = l)$, and $\gamma_{jl} = P(X' = j|H' = l)$. In words, α_{ik} is the fraction of traffic at entrance k from source i, β_{kl} is the probability of going from entrance k to exit l on the highway, and γ_{jl} is the fraction of traffic at exit l that proceed to destination j. The double sum in the expression is over all highway entrances and exits. Note that the traffic model is probabilistic, and in general allows for more than one highway route from source to destination. We further impose a constraint equating the onramp and offramp traffic distributions:

$$\gamma_{jl} = \alpha_{jl}.$$

Thus the fraction of traffic at exit l which continue to destination j is equal to the fraction of traffic at entrance l which originate from j. The model parameters are specified by $\alpha(x|h) = P(x|h)$ and $\beta(h,h') = P(h,h')$, which specify respectively the onramp/offramp traffic distribution, and highway traffic between the entrances and exits. Let the total amount of observed traffic be $N = \sum_{i,j} n_{ij}$, and let $\tilde{p}_{ij} = n_{ij}/N$ be the observed empirical joint distribution $\tilde{p}(x = i, x' = j)$. The log-likelihood function is given by

$$\mathcal{L} = N \sum_{x,x'} \tilde{p}(x,x') \log[\sum_{h,h'} \alpha(x|h)\beta(h,h')\alpha(x'|h')].$$

Maximizing the likelihood of the observed source-destination traffic counts is equivalent to minimizing the following Kullback-Leibler divergence:

$$\mathcal{D}(\tilde{p}(x,x') \parallel \sum_{h,h'} \alpha(x|h)\beta(h,h')\alpha(x'|h')).$$

The EM algorithm gives the following update equations
E-step

$$q(h,h'|x,x') = \frac{p(x,x',h,h')}{\sum_{hh'} p(x,x',h,h')}$$

where $p(x,x',h,h') = \alpha(x|h)\beta(h,h')\alpha(x'|h')$.
M-step

$$\alpha(x|h) = \frac{\tilde{p}(X = x, H = h) + \tilde{p}(X' = x, H' = h)}{\tilde{p}(H = h) + \tilde{p}(H' = h)}.$$

$$\beta(h,h') = \tilde{p}_{hh'},$$

where $\tilde{p}_{xh}$, $\tilde{p}_{x'h'}$, $\tilde{p}_h$, $\tilde{p}_{h'}$, and $\tilde{p}_{hh'}$ are the corresponding marginals of $\tilde{p}_{xx'}q(h,h'|x,x')$.

Representing the model parameters α and β as matrices, the highway traffic model seeks an approximation of the empirical traffic distribution $\tilde{p}$ by minimizing

$$\mathcal{D}(\tilde{p} \parallel \alpha\beta\alpha^t).$$

In comparison, a traffic model with the same structure as pLSA/NMF [4][7][8] seeks to minimize

$$\mathcal{D}(\tilde{p} \parallel AB).$$

The traffic interpretation of this model, which will be referred to as the “hub” traffic model, consists of an onramp distribution to the hubs from the sources, the hub distributions, and the offramp distributions from hubs to destinations. The highway model assumes more structure in the traffic data, and is a constrained version of the hub model. In particular, a highway model can always be represented as a hub model by equating corresponding terms in $(\alpha\beta)(\alpha^t) = (A)(B)$, effectively folding in the highway traffic between entrances and exits into the onramp traffic distribution specified by A. This comes at the cost of reduced sparseness of the onramp traffic distribution, and an increase in complexity of the hub model. Without equating onramp to offramp traffic in the highway model, the highway traffic has extra degrees of freedom since we can always write $\alpha\beta\gamma = (\alpha\beta)(I)(\gamma)$. Here the onramp traffic incorporates the highway traffic, and now there is no cross-traffic between entrances and exits. By equating onramp to offramp traffic, these degrees of freedom are removed in the highway traffic term β.

The highway and hub traffic models differ in complexity, sparseness and structure. In Section 3.1, the highway and hub traffic models will be compared using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores, as well as predictive test-set log-likelihoods. In Section 3.2, we demonstrate the extraction of a highway backbone structure in a random walk traffic matrix.

3. Numerical Experiments

3.1 Synthetic graph analysis

We start with a simple example analysis which elucidates the highway traffic model’s ability to find communities and their interrelations. A simple synthetic graph consisting traffic between 12 nodes is depicted on the left in Figure 1. Directed edges correspond to one traffic count in the given direction, whereas undirected edges represent a traffic count in both directions. The empirical joint source-destination distribution for the graph has exact decompositions according to both the highway and hub traffic models, with zero KL-divergences. Thus, the comparison here

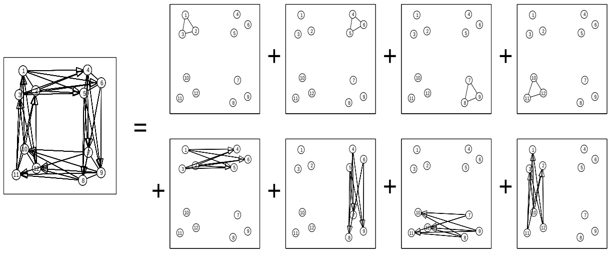

Figure 1. Synthetic graph decomposition based on the highway traffic model. The decomposition consist of four subgraphs of tightly knit communities and four subgraphs of relations between the communities.

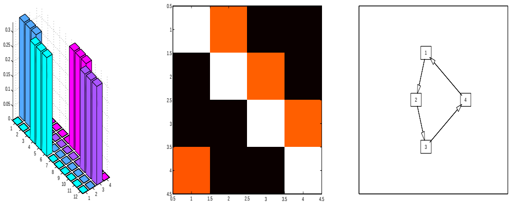

Figure 2. The highway model's on-ramp/offramp distribution (left), highway traffic β (center), and highway traffic visualization (right).

is in terms of the structure each model extracts from the data. For the highway model, the EM algorithm described above is run for 100 iterations starting from random initializations for $\alpha(x|h)$ and $\beta(h, h')$. The algorithm often finds the exact decomposition as shown in the figure. The exact decomposition of the graph consists of $k = 4$ fully connected communities consisting of 3 nodes each, given by $\alpha(x|h = i)\beta(h = i, h' = i)\alpha(x'|h' = i)$. These are depicted in the top four subgraphs on the right in Figure 1. In addition, the relations between the communities, as given by $\alpha(x|h = i)\beta(h = i, h' = j)\alpha(x'|h' = j)$, is depicted on in the bottom four subgraphs.

In Figure 2, the onramp/offramp distribution parameter α, and the highway traffic parameter β are displayed. In addition, a binary representation of the graph's highway backbone structure is visualized by thresholding β. For comparison, we fit a hub traffic model to the data. The corresponding graph decomposition if shown in Figure 3. The hub model all traffic within communities together with all outbound traffic from that community. The hub model essentially incorporated the highway traffic distribution into the offramp distribution.

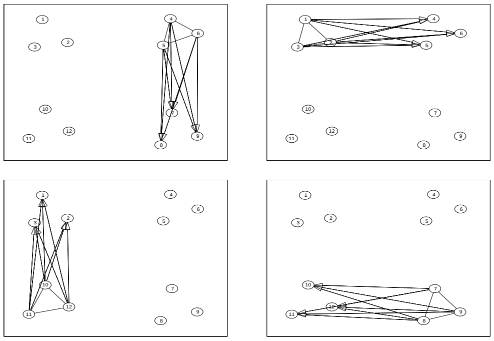

Figure 3. Synthetic graph decomposition based on the hub traffic model.

The highway model's analysis of this simple traffic graph successfully captures the tightly knit communities and their interrelations. In addition, the highway traffic matrix $\beta(h, h')$ describes the highway backbone traffic structure in the data.

3.2. Autonomous system connectivity graph

We analyzed simulated internet traffic data based on an undirected connectivity graph between Autonomous Systems (AS). The connectivity graph consists of AS paths in BGP routing tables collected by the server *route-views.oregon-ix.net*. This data is the basis of the power-law analysis in [11], and is publicly available at *http://topology.eecs.umich.edu/data.html*. After trimming out nodes with no edges, we are left with an undirected binary AS connectivity graph with 13233 interconnected AS nodes.

We compared the highway traffic model to the hub traffic model normalizing for the complexity differences of the two models. With k latent hub states in the hub model, and n sources/destinations, the hub model has $[2k(n-1)+k-1]$ parameters. In contrast, the highway model with the same number k or entrances/exits contains only $[k(n-1)+k^2-1]$ parameters. We compared the two models using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and predictive test-set log-likelihoods. The simulated traffic data was constructed as follows. For the training set, we performed 100000 *single* random walk steps on the connectivity graph. The 100000 source nodes were sampled in accordance with the stationary distribution of the random walk on the connectivity graph. Traffic from source nodes are assumed to follow each of the edge paths with equal probability. Since multinomial mixture models can be prone to over-fitting problems, we added a single pseudo-count traffic for each edge in the connectivity graph. If traffic from a source to a destination is not observed in the test set, but appears in the training set, the traffic models may assign zero probability to the test set likelihood. Early stopping will effectively stop parameter updates if an update assigns zero probability to a traffic path that appears in the test set. An additional inverse annealing (heating) is used in [8] to smooth multinomial parameters and prevent sparseness. For the test set, 20000 single random walk steps were simulated.

In Table 1 the AIC and BIC scores for the highway and hub models are tabulated for a number of different k values. For each model and each k, 10 EM runs with random parameter initializations are performed. Scores for the best respective runs are reported in the table. The highway traffic model has significantly better (lower) AIC and BIC scores than the hub traffic model.

values $\times 10^6$	k=26	k=51	k=100	k=197
Highway AIC	**4.90**	**5.41**	**6.59**	**9.07**
Hub AIC	5.56	6.72	9.16	14.15
Highway BIC	**8.34**	**12.2**	**19.9**	**35.0**
Hub BIC	12.4	20.2	35.5	66.1

Table 1. AIC and BIC scores for the highway and hub models.

We also compared predictive test-set log-likelihoods for a highway model and hub model with comparable degrees of freedom. Comparing the $k = 51$ highway model with 677432 parameters with the $k = 26$ hub model with 688089 parameters, the best test set log-likelihoods ($\times 10^5$) were -2.64 and -2.74 for the highway and hub models respectively. Comparing the $k = 100$ highway model with 1333199 parameters with the $k = 51$ hub model with 1349714 parameters, the best test set log-likelihood($\times 10^5$) were -2.55 and -2.63 respectively. Finally, the $k = 197$ highway model with 2645512 parameters and the $k = 100$ hub model with 2646499 parameters had best test set log-likelihoods($\times 10^5$) of -2.46 and -2.54 respectively. In all three comparisons, the highway model had slightly fewer parameters, but significantly higher (less negative) predictive test set log-likelihoods.

3.3. Random walk traffic on computer skills graph

Aside from complexity and sparseness considerations, the highway model extracts underlying backbone traffic structure which clustering models like the hub model does not. We analyzed a smaller, more easily interpretable data set to try to find communities and the relationships between the communities. A computer jobs description data set, provided courtesy of Prof. Richard Martin and the IT consulting firm Comrise was analyzed. The raw data consists of a collection of computer job descriptions, each of which contain a subset of 159 computer skills the hiring manager considered important for the job. The most frequently occurring skills keywords in the job descriptions are “unix”, “pc(ibm)”, “windows95”, “windowsnt”, “c” and “oracle”. Entries along the diagonal of the co-occurrence matrix contain the number of times each skill occurred over all the job descriptions. The elements of this matrix is interpreted as a the amount of (co-occurrent) traffic between pairs of job skills. This interpretation is equivalent to the normalization used in the random walk view of segmentation [9]. From the co-occurrent traffic information on the computer skills graph, we seek to extract out underlying computer skill topic communities, and the underlying backbone connectivity structure between the topic communities.

A visualization of the computer skills traffic graph is shown in Figure 4(a) using the *GraphViz* [12] spring model graph layout program from *AT&T Labs-Research*. Only edge connections with average transition probability greater than $.085$ are shown. Even though the graph is not very large with 159 nodes, the visualization is not easily readable, and only provides vague clues to relationships between various skills.

From the job skills co-occurrence table the observed empirical joint distribution $\tilde{p}_{xx'}$ is constructed. The EM algorithm is used to find the maximum likelihood estimators for the conditional $\alpha(x|h)$ and the joint $\beta(h, h')$.

Since the onramp and offramp traffic distributions are equal in the highway model, we will simply refer to the offramp traffic. The offramp traffic distribution from a few exits are tabulated in Table 2 This specifies the fraction of traffic at the specified exit which flow to each destination node. The destination computer skill with the largest traffic fraction is used as the label for the exit. The five top skills, ranked in descending order of their conditional probabilities are shown for each exit. From Table 2, we see a **UNIX** skills community containing *Unix*, *C* and *C++*, and a **SUNOS** operating systems community containing *SunOS*, *solaris* and *sunsparc*, and an **HP** cluster with *HP*, *HP-UX*. The model also identified skills groups affiliated with Microsoft, containing skills *PC(IBM), Windows95, MSoffice, MSproject* and *dos*, and a *Java* group (not tabulated) containing *javascript, perl, cgi* and *html*.

In addition to the communities of related computer skills, the model also extracts out the relationships between the communities. In Figure 4(b), we used *GraphViz* [12] to visualize the underlying highway traffic between entrance and exit hubs as defined by $\beta(h, h')$. This backbone traffic structure in the source-destination traffic data is visualized by thresholding $\beta(h, h')$ into a binary adjacency matrix. We emphasize that this is only for visualization purposes; the model contains more information than is visualized. From the highway traffic graph, we see tightly coupled highway

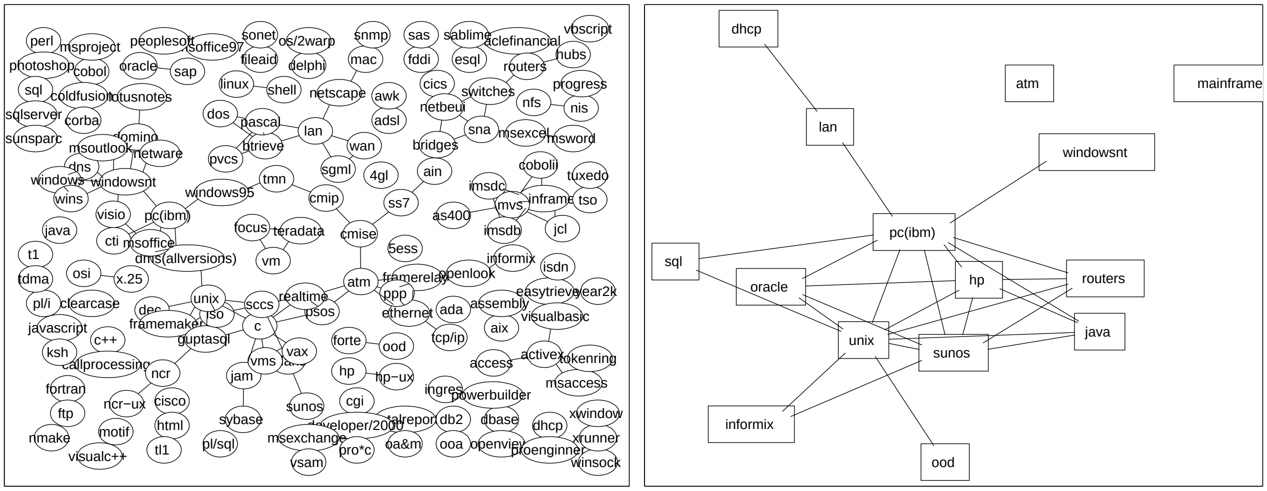

Figure 4. (a) Graph layout of the computer skills traffic graph using *GraphViz's* spring model layout algorithm. (b) Highway traffic visualization - each node in this graph is a highway entrance or exit hub, and corresponds to a computer skills community. Onramp and offramp distributions to sources and destinations are tabulated in Table 2.

Unix	.156	SunOS	.174	HP	.181	pc/ibm	.214
c	.154	solari	.169	hp-ux	.146	win95	.184
c++	.123	tuxedo	.016	tcpip	.076	msoff	.146
syb	.050	sunspa	.009	nis	.008	mspro	.050
jam	.004	oa&m	.007	nfs	.007	dos	.027

Table 2. Onramp/offramp traffic distribution for highway traffic model. The skills with highest traffic fraction to/from the latent states are listed in the first row, and used to label the clusters in Figure 4(b). Each column represents an entrance/exit hub. Fractions of traffic to/from each skill is listed next to the skill name.

traffic between the **Unix, SunOS, HP** communities, as well as the **Java** and **SunOS** communities. The highway traffic model successfully finds computer skills communities as well as the relationships between the communities.

4. Symmetric Hub Traffic Model

The highway traffic model assumes that the traffic is generated from an underlying highway traffic distribution, onramp traffic distributions from sources to highway entrances, and an identically distributed offramp distribution from highway exits to destinations. In contrast, the hub traffic model only has onramp and offramp distributions, and no analog of highway traffic. As discussed in Section 2, the hub model contains the highway model as a special case, where the composition of the onramp and highway traffic is subsumed into a single onramp traffic distribution for the hub model. Using the hub model to describe traffic data comes at the complexity cost of roughly double (for $n >> k$) the number of parameters as the highway model. An even more restrictive traffic model can be defined by equating onramp and offramp traffic distributions in the hub model. A model with this structure was first investigated in [14]. This "symmetric hub" model can be obtained from the highway model by constraining $\beta(h, h')$ to be diagonal.

We assume the traffic flow in the empirical joint distribution $\tilde{p}(x, x')$ is symmetric with respect to x and x'. This implies that the transition matrix $\tilde{p}(x'|x)$ is consistent with a reversible random walk. Instead of minimizing the Kullback-Leibler divergence for the highway traffic model:

$$\mathcal{D}(\tilde{p}(x, x') \parallel \sum_{h,h'} \alpha(x|h)\beta(h, h')\alpha(x'|h')),$$

the symmetric hub model minimizes

$$\mathcal{D}(\tilde{p}(x, x') \parallel \sum_{h,h'} \alpha(x|h)\beta(h)\alpha(x'|h')).$$

The EM algorithm for this model results in the iterations:

E-step

$$p(h|x,x') = \frac{\alpha(x|h)\alpha(x'|h)\beta(h)}{\sum_h \alpha(x|h)\alpha(x'|h)\beta(h)}$$

M-step

$$\beta(h) = \sum_{x',x} p(h|x,x')\tilde{p}(x',x),$$

$$\alpha(x|h) = \sum_{x'} \frac{p(h|x,x')\tilde{p}(x',x)}{\beta(h)}.$$

This symmetric hub model is a constrained version of both the hub model, and the highway model as follows. First, the symmetric hub model can be seen as a hub model with identically distributed onramp and offramp distributions. Second, the symmetric hub model is a highway model with the constraint of no traffic between the entrance/exit hubs.

This model does not directly capture relationships between communities since it does not directly model traffic between hub communities. However, it can describe traffic between hubs after *two* time steps of the empirical source-destination transition. An inter-hub traffic can be specified by combining the offramp traffic from hubs to destinations during the first time step, with the onramp traffic from destinations back to the hubs during the second time step. This is computed as follows:

$$p(h,h') = \sum_x \alpha(x|h)\beta(h)\alpha(x|h').$$

This *2-step* inter-hub traffic layout for the computer skills data is shown in Figure 5, with the onramp/offramp distributions tabulated in Tables 3 and 4. The skills topic communities successfully combine unix groups, windows groups and programming language groups, while the inter-hub traffic successfully represents the relationships between topic groups.

SunOS	.172	HP	.168	C++	.147
unix	.145	hp-ux	.139	C	.067
solaris	.142	tcp/ip	.080	ood	.060
tuxedo	.016	nis	.009	nmake	.031
sunsparc	.010	nfs	.008	dec	.007

Table 3. Onramp/offramp traffic distribution for highway traffic model the hub traffic model in Figure 5.

5. Highway traffic model properties

The source-destination traffic data analyzed in this paper either directly translated into symmetrical empirical joint distributions (computer skills), or were simulated from a reversible random walk (AS traffic). The highway traffic model on the other hand can in general describe non-symmetric traffic data. Looking at the model in more detail, the equating of the onramp with the offramp traffic distribution in the model results in the following conditional distribution within each community:

$$p(x,x'|h=h') = \alpha(x|h)\alpha(x'|h).$$

This is the highway model's predicted probability of transiting from source x to highway entrance h, and immediately exiting to destination x'. This conditional distribution matrix has rank 1 and satisfies the detailed balance condition. Thus, within each community, the random walk traffic is symmetric, and one can show that $\pi_h(x) = \alpha(x|k)$ is simply the stationary distribution of the random walk within each community. Even though the random walk within each community is reversible, the highway model can model non-reversible traffic depending on the highway

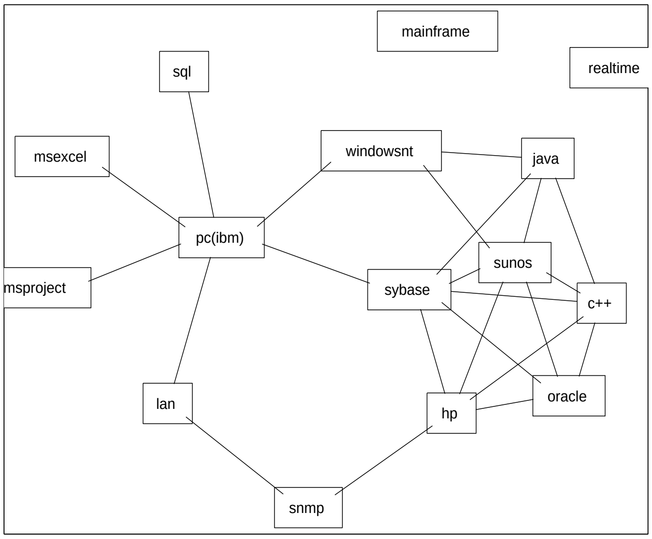

Figure 5. Symmetric hub model inter-hub traffic graph.

PC(IBM)	.235	windowsnt	.147	msexcel	.101
win95	.197	dos	.068	msword	.074
msoffice	.123	visualc++	.060	outlook	.011
lot-notes	.014	visualbasic	.031	visio	.011
mac	.002	vbscript	.007	isdn	.009

Table 4. Continuation of onramp/offramp traffic distribution for Figure 5.

traffic distribution $\beta(h, h')$.

If the highway traffic $\beta(h, h')$ between communities is symmetric with respect to source community (entrance) and destination community (exit), thereby satisfying the detailed balance condition, then the highway model describes symmetric source-destination traffic. One can verify that if the empirical traffic distribution is symmetric, and the highway traffic distribution $\beta(h, h')$ is initialized symmetric, then it will remain symmetric under all subsequent updates under the EM algorithm. Thus reversible source-destination traffic will be modeled with a reversible highway traffic model.

The traffic model approximates the empirical traffic flow in a maximum likelihood or minimum KL-divergence sense. For example, an approximation of the source traffic distribution can be obtained as follows. Let the source distribution of the highway traffic model be $\pi(x) = \sum_{h,h'} \alpha(x|h)\beta(h, h')$. Using Pinsker's inequality [13] we can bound the total variation distance between the empirical source distribution $\tilde{p}(x)$ and the source distribution of the highway model $\pi(x)$

$$\begin{aligned} & \sum_{x,x'} |\tilde{p}(x) - \pi(x)| \\ \le \quad & \sqrt{2\mathcal{D}(\tilde{p}(x) \parallel \pi(x))} \\ \le \quad & \sqrt{2\mathcal{D}(\tilde{p}(x,x') \parallel \sum_{h,h'} \alpha(x|h)\beta(h,h')\alpha(x'|h'))}. \end{aligned}$$

Similarly, the highway model can approximate any empirical traffic flow from a source set of nodes to a destination set, with the KL-divergence providing a bound on the approximation error.

6. Discussion

The symmetric hub model, highway model, and hub model are constructed with various structures and associated complexities in the multinomial mixture. This is analogous to controlling the covariance structure in Gaussian distributions, from spherical Gaussian, to Graphical Gaussian models and Factor Analysis, to the full Gaussian with arbitrary covariance structure.

We compared the highway and hub traffic models using the Akaike Information Criterion and also test set log-likelihood for the ASP data set. The suitability of each traffic model clearly depend on the underlying structure of the empirical source-destination traffic. Consider for example, source-destination *car* traffic. One could conceivably build a road system based on the hub traffic model, with onramps from origins to k underlying hubs, and offramps to the destinations. This could capture a highway traffic model distribution with k cities, and highway traffic between them. However, the added complexity of the hub model comes at the significant cost of building non-sparse onramps or offramps. In the extreme limit, one could build roads between all origins and destinations with empirical traffic counts. The benefit of the highway model is in the aggregation of traffic flow along an underlying highway infrastructure. An important consideration in comparing the models should be sparseness of the resulting traffic model. There will in general be domain specific sparseness related cost functions to consider.

Hub models with the same probabilistic structure as pLSA/NMF have been applied in the information retrieval setting to decompose document-word matrices [7][8] and document-citation matrices [15]. In those settings, pLSA does not provide a probabilistic generative model, and is not able to cleanly assign predictive probabilities to new documents. Latent Dirichlet Allocation [16] improves on pLSA by providing a proper generative model. In the source-destination traffic setting we consider, the sources/destinations constitute a fixed set, and the traffic models properly defines probabilities for new traffic between the sources and destinations. The traffic models are properly defined probabilistic models of source destination traffic. Over-fitting however, can be a problem. Specifically, if traffic from a source to a destination is not observed in the test set, but appears in the training set, the traffic models may assign zero probability to the test set likelihood. One can use smoothing or incorporate priors over the multinomial parameters.

In summary, the highway model extracts out communities and relational information in the form of highway traffic between the communities. It is related to spectral clustering algorithms where the interest is in finding communities of nodes with minimal traffic between the communities [9][10]. The highway traffic model extends the framework of minimizing traffic flow between communities and provides a low rank highway based approximation to the empirical source-destination traffic. In the relational data research field, models have been investigated in the context of binary link detection [17], binary relational modeling [18], and in a supervised learning context for link prediction. [19]. We are pursuing extensions of the highway traffic model to address the selection of the number of highway entrances/exits, as well as traffic models with highways and freeways. The highway model can also be extended from an unsupervised to a semi-supervised setting with some observations of highway and onramp/offramp traffic counts.

7. Acknowledgments

This work was supported by the National Science Foundation (NSF Grant DMS-0312275).

8. References

[1] Vardi, Y. (1996) Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *Journal of the American Statistical Association*, Vol.91, No. 433, pp365-377.

[2] Cao, J., Davis, D., Wiel, S.V. and Yu, B. (2000) Time-Varying Network Tomography. *Journal of the American Statistical Association*, Vol.95, No. 452, pp.1063-1075.

[3] Everitt, B. (1984). An Introduction to Latent Variable Models. London: Chapman & Hall.

[4] Saul, L. and Pereira, F. (1997) Aggregate and mixed-order Markov models for statistical language processing. In *Proceedings of the second conference on empirical methods in natural language processing*, 81-89.

[5] Brown,P., Della Pietra, V., deSouza,P., Lai. J. (1992). Class-based n-gram models of natural language. Computational Linguistic, 18:467-479, 1992.

[6] Pereira, F., Tishby, N., and Lee, L. (1993) Distributional clustering of English words. In *Proceedings of the ACL*, pp183-190, 1993.

[7] Lee, D. and Seung, S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(675), 788-791.

[8] Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196.

[9] Meila, M. and Shi, J. (2001) A random walks view of spectral segmentation. in *Proc. International Workshop on AI and Statistics (AISTATS), 2001*

[10] Ng, A., Jordan, M, and Weiss, Y. (2001) On spectral clustering: analysis and an algorithm. In *NIPS 14* 2001.

[11] Qien, C., Chang, H., Govindan, R., Jamin, S., Shenker, S. and Willinger, W. (1992) The Origin of Power Laws in Internet Topologies Revisited, *Proc. of IEEE Infocom, 2002*

[12] GraphViz, AT&T Labs-Research.
URL: *http://www.research.att.com/sw/tools/graphviz*

[13] Pinsker, M. (1960) Information and information stability of random variables and processes. Moskow: Izv. Akad. Nauk, 1960.

[14] Lin, J. (2003) Reduced-Rank Approximations of Transition Matrices, in C. M. Bishop and B. J. Frey (eds), Proceedings of AI-Statistics'2003, Jan 3-6, 2003, Key West, FL.

[15] Cohn, D. and Hofmann, T. (2001) The Missing Link A Probabilistic Model of Document Content and Hypertext Connectivity. In Advances in Neural Information Processing Systems 13

[16] Blei, D., Ng, A. and Jordan, M. (2003) Latent Dirichlet Allocation, The Journal of Machine Learning Research, vol 3, pp.993-1022

[17]Schneider, J., Kubica, J., Moore, A. and Yang, Y. (2002) Stochastic link and group detection. In *IJCAI* 2002.

[18] Kemp, C., Griffiths, T. and Tenenbaum, J. (2004) Discovering latent classes in relational data. AI Memo 2004-019, M.I.T. AI Laboratory.

[19] Taskar, B.,Wong, M.F., Abbeel, P. and Koller, D. (2003) Link Prediction in Relational Data. Neural Information Processing Systems Conference (NIPS03), Vancouver, Canada, December 2003.