Experimenting with Drugs (and Topic Models): Multi-Dimensional Exploration of Recreational Drug Discussions

Michael J. Paul and **Mark Dredze**
Center for Language and Speech Processing
Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD 21218

Abstract

Clinical research of new recreational drugs and trends requires mining current information from non-traditional text sources. In this work we support such research through the use of a multi-dimensional latent text model – *factorial LDA* – that captures orthogonal factors of corpora, creating structured output for researchers to better understand the contents of a corpus. Since a purely unsupervised model is unlikely to discover specific factors of interest to clinical researchers, we modify the structure of factorial LDA to incorporate prior knowledge, including the use of of observed variables, informative priors and background components. The resulting model learns factors that correspond to drug type, delivery method (smoking, injection, etc.), and aspect (chemistry, culture, effects, health, usage). We demonstrate that the improved model yields better quantitative and more interpretable results.

Introduction

Topic models aid exploration of the main thematic elements of large text corpora by producing a high level semantic view (Blei, Ng, and Jordan 2003; Eisenstein et al. 2012). Topic models have been used for understanding the contents of a corpus and identifying interesting aspects of a collection for more in-depth analysis (Talley et al. 2011; Mimno 2011).

We consider a large collection of Web discussion forums about recreational drug usage: such data are becoming a common information source of clinical studies of new drugs (Corazza et al. 2011; Hill and Thomas 2011; Schifano et al. 2006; Corazza et al. 2012). While a topic analysis may identify different drugs, it is only one of many ways to analyze the corpus. In fact, there are specific factors of interest to medical researchers, such as different drug delivery methods (oral, injection, smoking, etc.) or aspects of drug usage (cultural settings, health ramifications, drug chemistry, etc.) We seek a model that jointly captures these factors, rather than modeling each in isolation. Automated discovery of these factors can aid in drug discovery and usage details, an improvement over the current approach of manual forum analysis.

Towards this goal we use *factorial LDA* (f-LDA), a recently introduced general framework for multi-dimensional text models that captures an arbitrary number of factors (Paul and Dredze 2012). While a standard topic model learns document specific topic distributions, f-LDA learns distributions over combinations of factors (e.g. drug, delivery and aspect) called tuples, e.g. (CANNABIS,SMOKING,EFFECTS). We use f-LDA to model three factors of **drug type**, **delivery method** and **aspect** by modifying the model to incorporate prior knowledge about these factors. We demonstrate that the resulting model captures factors of interest to the user, as demonstrated through improved quantitative results and model interpretability.

Factor	Components
Drug	ALCOHOL AMPHETAMINES ANTIDEPRESSANTS BETA-KETONES CANNABINOIDS CANNABIS COCAINE DMT DOWNERS DXM ECSTASY GHB HERBAL ECSTASY KETAMINE KRATOM LSA SEEDS LSD MAGIC MUSHROOMS NOOTROPICS OPIOIDS PEYOTE PHENETHYLAMINES SALVIA TOBACCO
Delivery	GENERAL INJECTION ORAL SMOKING INSUFFLATION (SNORTING)
Aspect	GENERAL CHEMISTRY (Pharmacology, TEK) CULTURE (Culture, Setting, Social, Spiritual) EFFECTS (Effects) HEALTH (Health, Overdose, Side effects) USAGE (Dose, Storing, Weight)

Table 1: The three factors of our model.

Tracking Drug Trends for Public Health

Recreational drug use imposes a significant burden on the health infrastructure of the United States and other countries. Accurate information on drugs, usage profiles and side effects are necessary for supporting a range of healthcare activities, such as drug addiction treatment programs, toxin diagnosis, prevention, safety awareness campaigns and public policy. These activities rely on up to date information on drug trends as substance popularity changes in response to legislative efforts and market trends. Hospitals and poison control centers among others must remain informed on the pharmacological and toxicological effects of new and popular drugs (Hill and Thomas 2011). Understanding usage patterns can inform outreach strategies (Bruneau et al. 2012).

A number of sources aid in studying drug trends. The most accurate information comes from speaking directly with users, e.g. focus groups (Reyes et al. 2012) or interviews (Hout and Bingham 2012). Alternative, less time consuming, methods include wastewater testing for known toxins (Wish et al. 2012; Zuccato et al. 2011), tracking ICD-10 codes from hospitals that correlate with toxicity (Shah, Wood, and Dargan 2011), or testing chemicals found in ER patients (Wood et al. 2011). While faster, these methods provide a skewed and incomplete picture.

While online drug discussions were first viewed as a dangerous information source on designer drugs for users (Wax 2002), researchers now recognize clinical value in this information (Corazza et al. 2011). Morgan, Snelson, and Elison-Bowers (2010) found drug images pervasive on popular social media websites, and some sites targeted for recreational drugs provide a detailed picture of drug use. Comprehensive reviews now include standard (PUBMED) and non-standard sources: media reports, government publications and drug user web forums (Hill and Thomas 2011). These forums are especially helpful for new drugs that arise as legal alternatives to banned drugs (Gallagher et al. 2012). The EU Psychonaut project focuses on categorizing recreational drug information from online forums (Schifano et al. 2006).

Consider an illustrative example from recent work by Corazza et al. (2012): the new drug methoxetamine, a ketamine derivative. Ketamine is a controlled substance and methoxetamine is a popular legal psychoactive alternative. However, methoxethamine has no clinical trials and thus little is known about its use, effects, or popularity. Corazza and colleagues turned to forums for information on the drug's effects and usage. A manual analysis of online materials, such as Drugs-Forum (discussed below) and YouTube videos advertising the drug, uncovered such details.

Organizing and understanding forums requires significant effort; manual analysis is time consuming. Instead, we propose automated tools for exploration and analysis of these data. Approaches based on supervised models, popular in surveillance, cannot capture new drugs of which researchers are unaware (Winstock and Mitcheson 2012). In fact, existing surveillance through traditional indicators (e.g. hospitals and law enforcement) fails to identify the emergence of new drug classes, such as mephedrone (Dunn et al. 2011).

Instead, we rely on unsupervised topic models, where the identification of thematic elements can uncover emerging trends. Topic models have been shown to be useful tools for studying public health in Web data such as Twitter (Paul and Dredze 2011). However, standard topic models cannot capture the diversity of factors of interest: drugs, delivery method, various aspects, etc. Instead, a multi-dimensional text model can simultaneously capture these different factors, providing a more informative understanding of the data. In this work, we modify f-LDA to incorporate prior knowledge to discover factors of interest in drug discussions.

Corpus: *Drugs-Forum*

Our data set is taken from `drugs-forum.com`, a site that has been active for more than ten years with over 100,000 members and more than one million monthly readers. The site is an information hub where people can freely discuss recreational drugs with psychoactive effects, ranging from coffee to heroin, hosting information and discussions on specific drugs, as well as drug-related politics, law, news, recovery and addiction. Site users are primarily drug users, but also include researchers, parents, officials, NGOs, lawyers, doctors, journalists and addiction specialists. The site has been used in clinical research (Corazza et al. 2012).

Discussion threads are organized into numerous forums, including drugs, the law, addiction, etc. Since our interest was learning about drug use, we focus on the drug forums. Each thread is assigned to a specific forum (drug) and each thread has a user-specified tag, which can indicate delivery method (e.g. "oral"), or general categories like "effects." We focused on a few tags of interest, shown in Table 1.

Multi-Dimensional Text Models

We begin by summarizing **factorial LDA (f-LDA)** (Paul and Dredze 2012), a multi-dimensional text model that jointly captures multiple orthogonal semantic *factors*.[1]

Consider a standard topic model (e.g. LDA (Blei, Ng, and Jordan 2003)) where the choice of topics corresponds to selecting entries in an array; document specific topic distributions are distributions over the array. In f-LDA, which captures K factors, we replace the flat array with a K-dimensional array; document specific distributions are over the K-dimensional array. Each dimension is called a *factor*, the specific choice for an entry along one factor a *component*, and the combination of a component from each factor forms a *tuple*, i.e. an entry in the K-dimensional array. In our application, the first factor will be drugs, the second delivery method, and the third aspect. An example tuple could be (CANNABIS,SMOKING,EFFECTS). In the same way that each topic is associated with a word distribution in LDA, each tuple is associated with a word distribution in f-LDA.

Formally, $\theta^{(d)}$ is a document specific distribution over a K-dimensional array, and each token is associated with a latent vector $\vec{z}$ of length K; we have K factors, each with Z_k components. The Cartesian product of the K factors forms a set of tuples and the vector $\vec{z}$ references K components to form a tuple $\vec{t} = (t_1, t_2, \ldots, t_K)$. Each entry in the array (i.e. each tuple) references a word distribution that is influenced by the associated components. In this model, a token is generated by first sampling an entire tuple $\vec{z}$ from the document specific $\theta^{(d)}$ and then the token is sampled from the tuple's corresponding word distribution $\phi_{\vec{z}}$. $\theta^{(d)}$ is drawn from the prior Dirichlet($\hat{\alpha}$).

Intuitively, tuples which share components should have word distributions which share words. The word distributions for the triples (CANNABIS,SMOKING,EFFECTS) and (CANNABIS,ORAL,CHEMISTRY) should both contain words about cannabis. f-LDA solves this by utilizing a structured word prior to encourage similar words to appear in each tuple with the same component. $\phi_{\vec{t}}$, the word distribution for tuple $\vec{t}$, has a Dirichlet($\hat{\omega}_w^{(\vec{t})}$) prior for word w, where $\hat{\omega}_w^{(\vec{t})}$ is a log-linear function of three parameter types: $\omega^{(B)}$, a

[1] Full details can be found in Paul and Dredze (2012).

corpus-wide precision parameter (the bias), $\omega_w^{(0)}$, the corpus specific bias for word w, and $\omega_{t_k w}^{(k)}$, the bias parameter for word w for component t_k of the kth factor – it is this last parameter which ties together tuples that share the component t_k. These parameters are combined as $\hat{\omega}_w^{(\vec{t})} \triangleq \exp\left(\omega^{(B)} + \omega_w^{(0)} + \sum_k \omega_{t_k w}^{(k)}\right)$, which forms the Dirichlet prior for $\vec{t}$'s word distribution.

Another problem that f-LDA addresses is the fact that many tuples will have little support in the data, and so the Cartesian product of factors should be sparse – the posterior "opts out" of some tuples. To handle this, the prior over tuples becomes $\theta \sim \text{Dirichlet}(\mathbf{B} \cdot \hat{\alpha})$, where $\cdot$ is the cell-wise product and $\mathbf{B}$ is a sparsity inducing K-dimensional array, where an entry $b_{\vec{t}}$ corresponds to tuple $\vec{t}$. The values of b are in $(0, 1)$, where values close to 1 or 0 represent whether a tuple is active or inactive. If $b_{\vec{t}}$ is close to 0, then θ in each document will have a very low prior probability of choosing $\vec{t}$. This allows the model to avoid learning word distributions for tuples that do not have support – for example, (CANNABIS,INJECTION,EFFECTS) does not appear in the data because cannabis is not injected.

Performance Enhanced Factorial LDA

We will use f-LDA to model three factors relating to drug usage. In addition to drug type, the two other factors are delivery method and general aspects of drug usage. First, researchers are interested in the method of drug delivery (injection, oral, smoking, etc.) as different delivery methods can yield different effects. Second, there are many aspects to drug usage, such as the cultural context, the effects, side effects, or health implications. Modeling these three factors yield tuples of the form (COCAINE,SNORTING,HEALTH) and (CANNABIS,SMOKING,CULTURE).

However, without any supervision, there is no way to ensure that the model will actually discover these three factors. In this section, we present a 3-dimensional f-LDA model augmented with prior knowledge, to encourage the model to learn the factors of interest.

Table 1 shows the different components of the three factors that we are going to model. The many sub-forums within our data are already categorized (sometimes hierarchically) by drug or drug class, which gives the components of the first factor (drug type). Because of this organization, we treat this factor as an *observed* variable during learning. Thus, messages from the "cannabis" forum will use tuples of the form (CANNABIS,*,*).

The delivery method and aspect factors are not observed, but we can still make use of side information to guide the model. Each discussion thread is tagged with exactly one label, such as "Snorting" or "Side effects," and these tags give us an incomplete set of labels for threads. A number of tags correspond directly to delivery method, and others are manually grouped into components for aspect: e.g. CULTURE (tags: Culture, Setting, Social, Spiritual).

We cannot simply use these tags for supervised learning because most documents are missing labels (only 30% of our corpus contains one of the labels in Table 1) and many messages discuss several components, not just the one implied by the tag. However, we can make use of the tags in a semi-supervised framework; specifically, we will use these tags to create *prior* probabilities over the word distributions for these components.

Tags and Word Priors We will now describe how to create these word priors based on tags. Assume for the moment we are given a general distribution over words in the corpus and a distribution over the words associated with a tag. This is formalized as a vector m of log-frequencies over the vocabulary for the whole corpus, and a vector $\eta_i^{(f)}$ of log-frequencies over the vocabulary for the ith component of factor f. If we had these values, we could use them to guide learning as prior knowledge over model parameters ω. While f-LDA assumes each ω is drawn from a 0-mean Gaussian, we alter the means of the appropriate ω parameters to use m and η:

$$\omega_w^{(0)} \sim \mathcal{N}(m_w, \sigma^2); \omega_{iw}^{(k)} \sim \mathcal{N}(\eta_{iw}^{(k)}, \sigma^2). \quad (1)$$

Recall that $\omega_w^{(0)}$ are corpus-wide bias parameters for each word and $\omega_{iw}^{(k)}$ are component specific parameters for each word. This yields a hierarchical prior in which η parameterizes the prior over ω, while ω parameterizes the prior over ϕ (the word distributions).

In additional to the components which come from the forum tags, we also add an extra component called GENERAL – with index 0 – to the second (delivery) and third (aspect) factors. General words that are not specific to individual components will fall to the general components – we set all $\eta_0^{(k)}$ to $\vec{0}$, so that there is no prior bias towards certain words.

Learning the Priors We have described the prior means m and η by assuming they are given to us. In reality, we must learn these from tagged messages. However, these parameters imply a latent division of responsibility for observed words: some are present because of the tag while others are general words in the corpus. These parameters must be estimated in a way that acknowledges this division.

We learn these parameters from the tagged messages using SAGE, which models words in a document as combinations of background and topic word distributions. Eisenstein, Ahmed, and Xing (2011) present SAGE models for Naive Bayes (one class per document), admixture models (one class per token), and admixture models where tokens come from multiple factors. We combine the first and third models, such that a document has multiple factors which are given as labels across the entire document – the drug type and the tag, which could correspond to a component of either the delivery or aspect factors. We posit the following model of text generation per document:

$$P(\text{word } w|\text{drug} = i, \text{factor} f = j) \quad (2)$$
$$= \frac{\exp(m_w + \eta_{iw}^{(1)} + \eta_{jw}^{(f)})}{\sum_{w'} \exp(m_{w'} + \eta_{iw'}^{(1)} + \eta_{jw'}^{(f)})}$$

As in SAGE, we fix m to be the observed vector of corpus log-frequencies over the vocabulary, which acts as an "overall" weight vector, while parameter estimation yields $\eta_i^{(f)}$,

the relative log-frequencies vector for the ith component of factor f. We learn the parameters by optimizing the model in (2) using gradient ascent. These parameters are then used as the mean of the Gaussian priors over ω.

We call this model augmented with prior knowledge Performance Enhanced Factorial LDA (pef-LDA).

Experiments

Our corpus consists of messages from `drugs-forum.com`. The site categorizes threads into many topics, including some on specific drugs, which are categorized hierarchically. We treat each top-level category as a drug type. While this works well for some drugs that are pharmacologically related and have similar effects, such as the opioids/opiates category which includes codeine, morphine, and heroin, it does not capture broader categories, such as the ethnobotanicals category, which includes a broad array of psychoactive plants as varied as the hallucinogenic peyote cactus and the opioid-like kratom leaf. In these cases, we instead treat the individual sub-category drugs separately, rather than lumping them into one top level category. We selected 24 popular drugs and from these forums we randomly selected a total of 100K messages (out of 409K). Each message in a thread was considered a separate document, and we only used documents with at least five word tokens after stopwords, punctuation and low frequency words were removed. This preprocessed data set contains an average of 45 tokens per document, with a total of 8.7K unique word types.

Model Learning All instances of pef-LDA are run with 5000 iterations of Gibbs sampling. We initialize the Gibbs sampler so that each token in a document is assigned to its label given by the tag, when available. In the absence of tags, we initialize tokens to the background components, so a large majority of tokens are initialized to the background. We initialize ω to its prior mean (Eq. 1).

We optimize the hyperparameters and sparsity array using gradient descent after each Gibbs sweep. We use a decreasing step size of $a/(t+1000)$, where t is the current iteration and a=10 for α and 1 for ω and the sparsity values. To learn the priors η, we run our version of SAGE for 100 iterations of gradient ascent, with a fixed step size of 0.1. The normal priors use σ^2=10.0 for α and 0.5 for ω.

Quantitative Validation We designed pef-LDA to capture particular factors in the data. To validate if pef-LDA captures these factors better than the out-of the-box f-LDA, we experimented with two predictive tasks on 25K held-out documents. First, we computed standard measurements of corpus perplexity. Second, we measured how well the model can predict the observed tags of threads, both in accuracy (how often the true tag was the model's most likely component) as well as the mean reciprocal rank (MRR) of the true tags.

Model	Perplexity	Accuracy	MRR
f-LDA	1765	14%	0.37
pef-LDA	1730	41%	0.62

Table 2: Quantitative comparison of f-LDA and pef-LDA.

CHEMISTRY	CULTURE	EFFECTS	HEALTH	USAGE
solvent	kids	feeling	symptoms	100mg
evaporate	police	visuals	depression	weighs
ethanol	weve	relaxed	severe	dose
tek	owner	felt	long-term	200mg
extraction	don	comedown	disorders	dosage
solvents	public	euphoria	syndrome	250mg
ethyl	war	feels	bodys	300mg

Table 3: The words with the highest learned ω values for five aspects, which affect the prior over the word distributions ϕ.

For f-LDA, we used a post-hoc greedy matching to determine which model components corresponded to which tag, based on the Jensen-Shannon divergence between each component's marginal distribution and the distribution defined by the prior. Table 2 shows that our model enhancements provide better predictive abilities.

What Does the Model Learn?

We've demonstrated quantitative benefits to pef-LDA and now focus on qualitative experiments, which reflect pef-LDA's ability to discover interesting drug patterns. pef-LDA learns word distributions for tuples combining the three factors. We present examples of the resulting tuples by selecting the top 6 words for each tuple (Table 4).

The structured output itself appears more informative than a flat list of topics to a researcher. This output breaks down words for each drug into delivery method and aspect. For example, the cocaine component distinguishes words between delivery methods: smoking (pipe, rock) vs. snorting (nose, powder), and aspects: chemistry (acetone, water) vs. health (addiction, brain). Additionally, the labels for drug, delivery method and aspect are not assigned manually, but taken from the prior; this both saves time and clarifies the output. The tuples clearly correspond to the labeled components.

An examination of even a small slice of output reveals several patterns of drug use, such as:

- **Cocaine:** The delivery methods reveal different types of cocaine. The SMOKING component has the words "crack" and "rock", while the SNORTING component has the words "coke", "powder" and "lines".
- **Cannabis:** The oral method includes words about marijuana brownies; the tuple (CANNABIS, ORAL, CHEMISTRY) contains words related to baking, such as "butter" and "milk", which are particular to this delivery method.
- The **culture** components reflect differences in the culture surrounding drugs. ECSTASY contains words related to raves and nightclubs, and OPIOIDS, which includes heroin, has words about addiction and street life ("money", "dealer", "junkie").
- The **health** components highlight health issues surrounding different types of drugs. COCAINE and OPIOIDS both include words about addiction, while CANNABIS includes words about mental health ("mental", "anxiety", "psychosis"). We also find health words that are specific to certain delivery methods: the tuple (COCAINE, SNORTING,

HEALTH) includes words about nose and sinus damage, and (CANNABIS, SMOKING, HEALTH) includes the words "cancer", "lung", and "lungs".

Additionally, Table 3 shows the top words (based on the prior hyperparameters ω) for some of the individual components, which illustrates how the priors for particular aspects cut across various tuples.

Conclusion and Future Work

To the best of our knowledge, this work represents one of the first investigations into using automated text processing techniques for analyzing documents from the recreational drug domain.[2] We have presented pef-LDA, an extension to factorial LDA tailored to a particular application and data set which was demonstrated to induce desired properties. This study thus lays out practical guidelines for customizing multi-dimensional text models for text analysis applications. In future work, we hope to extend pef-LDA to model finer-grained drug types in the hope of discovering lesser-known and new drugs. We also plan to use the output of this model to perform specific analyses of drug use, such as drug trends over time and usage variation across demographic groups.

Acknowledgements We are grateful to the anonymous reviewers for helpful feedback. The first author was supported by an NSF Graduate Research Fellowship.

References

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *JMLR*.

Bruneau, J.; Roy, É.; Arruda, N.; Zang, G.; and Jutras-Aswad, D. 2012. The rising prevalence of prescription opioid injection and its association with hepatitis C incidence among street-drug users. *Addiction* 107(7):1318–1327.

Corazza, O.; Schifano, F.; Farre, M.; Deluca, P.; Davey, Z.; Drummond, C.; Torrens, M.; Demetrovics, Z.; Di Furia, L.; Flesland, L.; et al. 2011. Designer drugs on the internet: a phenomenon out-of-control? the emergence of hallucinogenic drug bromo-dragonfly. *Current Clinical Pharmacology* 6(2):125–129.

Corazza, O.; Schifano, F.; Simonato, P.; Fergus, S.; Assi, S.; Stair, J.; Corkery, J.; Trincas, G.; Deluca, P.; Davey, Z.; Blaszko, U.; Demetrovics, Z.; Moskalewicz, J.; Enea, A.; di Melchiorre, G.; Mervo, B.; di Furia, L.; Farre, M.; Flesland, L.; Pasinetti, M.; Pezzolesi, C.; Pisarska, A.; Shapiro, H.; Siemann, H.; Skutle, A.; Enea, A.; di Melchiorre, G.; Sferrazza, E.; Torrens, M.; van der Kreeft, P.; Zummo, D.; and Scherbaum, N. 2012. Phenomenon of new drugs on the internet: the case of ketamine derivative methoxetamine. *Human Psychopharmacology: Clinical and Experimental* 27(2):145–149.

Coyle, J. R.; Presti, D. E.; and Baggott, M. J. 2012. Quantitative analysis of narrative reports of psychedelic drugs. *arXiv* 1206:0312.

Dunn, M.; Bruno, R.; Burns, L.; and Roxburgh, A. 2011. Effectiveness of and challenges faced by surveillance systems. *Drug Testing and Analysis* 3(9):635–641.

[2]Very recent work has investigated the use of supervised machine learning to classify recreational drugs from narratives on the Web (Coyle, Presti, and Baggott 2012).

Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse additive generative models of text. In *ICML*.

Eisenstein, J.; Chau, D. H. P.; Kittur, A.; and Xing, E. P. 2012. Topicviz: Semantic navigation of document collections. In *CHI Work-in-Progress Paper*.

Gallagher, C. T.; Assi, S.; Stair, J. L.; Fergus, S.; Corazza, O.; Corkery, J. M.; and Schifano, F. 2012. 5,6-methylenedioxy-2-aminoindane: from laboratory curiosity to 'legal high'. *Human Psychopharmacology: Clinical and Experimental* 27(2):106–112.

Hill, S. L., and Thomas, S. H. L. 2011. Clinical toxicology of newer recreational drugs. *Clinical Toxicology* 49(8):705–719.

Hout, M. C. V., and Bingham, T. 2012. Costly turn on: Patterns of use and perceived consequences of mephedrone based head shop products amongst Irish injectors. *International Journal of Drug Policy* 23(3):188–197.

Mimno, D. 2011. Reconstructing Pompeian households. In *UAI*.

Morgan, E. M.; Snelson, C.; and Elison-Bowers, P. 2010. Image and video disclosure of substance use on social media websites. *Computers in Human Behavior* 26(6):1405 – 1411.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing Twitter for public health. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Paul, M. J., and Dredze, M. 2012. Factorial LDA: Sparse multi-dimensional text models. In *Neural Information Processing Systems (NIPS)*.

Reyes, J.; Negrón, J.; Colón, H.; Padilla, A.; Millán, M.; Matos, T.; and Robles, R. 2012. The emerging of xylazine as a new drug of abuse and its health consequences among drug users in Puerto Rico. *Journal of Urban Health* 1–8.

Schifano, F.; Deluca, P.; Baldacchino, A.; Peltoniemi, T.; Scherbaum, N.; Torrens, M.; Farrõ, M.; Flores, I.; Rossi, M.; Eastwood, D.; Guionnet, C.; Rawaf, S.; Agosti, L.; Furia, L. D.; Brigada, R.; Majava, A.; Siemann, H.; Leoni, M.; Tomasin, A.; Rovetto, F.; and Ghodse, A. H. 2006. Drugs on the web; the psychonaut 2002 eu project. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 30(4):640 – 646.

Shah, A.; Wood, D.; and Dargan, P. 2011. Survey of ICD-10 coding of hospital admissions in the UK due to recreational drug toxicity. *QJM* 104(9):779–784.

Talley, E.; Newman, D.; II, B. H.; Wallach, H.; Burns, G.; Leenders, M.; and McCallum, A. 2011. A database of National Institutes of Health (NIH) research using machine learned categories and graphically clustered grant awards. *Nature Methods*.

Wax, P. 2002. Just a click away: recreational drug web sites on the internet. *Pediatrics* 109(6):e96–e96.

Winstock, A. R., and Mitcheson, L. 2012. New recreational drugs and the primary care approach to patients who use them. *BMJ* 344.

Wish, E. D.; Artigiani, E.; Billing, A.; Hauser, W.; Hemberg, J.; Shiplet, M.; and DuPont, R. L. 2012. The emerging buprenorphine epidemic in the United States. *Journal of Addictive Diseases* 31(1):3–7.

Wood, D. M.; Panayi, P.; Davies, S.; Huggett, D.; Collignon, U.; Ramsey, J.; Button, J.; Holt, D. W.; and Dargan, P. I. 2011. Analysis of recreational drug samples obtained from patients presenting to a busy inner-city emergency department: a pilot study adding to knowledge on local recreational drug use. *Emergency Medicine Journal* 28(1):11–13.

Zuccato, E.; Castiglioni, S.; Tettamanti, M.; Olandese, R.; Bagnati, R.; Melis, M.; and Fanelli, R. 2011. Changes in illicit drug consumption patterns in 2009 detected by wastewater analysis. *Drug and Alcohol Dependence* 118(2-3):464 – 469.

Aspect

Delivery Method	GENERAL	CHEMISTRY	CULTURE	EFFECTS	HEALTH	USAGE
	CANNABIS					
ORAL	weed	butter	friend	trip	sleep	time
	high	oil	went	experience	cannabis	pot
	eat	heat	night	lsd	dreams	hours
	eating	water	friends	hallucinations	memory	gram
	brownies	milk	home	psychedelic	effects	half
	work	mix	room	intense	experience	grams
SMOKING	tobacco	pipe	said	time	smoking	smoke
	joint	glass	marijuana	smoked	marijuana	bong
	weed	bowl	drug	weed	smoke	hit
	joints	water	police	felt	cannabis	bowl
	smoke	bottle	law	first	cancer	hits
	roll	hole	store	high	cause	smoking
	COCAINE					
GENERAL	dont	acetone	people	coke	cocaine	time
	know	wash	life	high	addiction	first
	think	water	friends	feel	drug	line
	coke	cocaine	drugs	cocaine	alcohol	lines
	people	product	time	meth	dopamine	gram
	want	pure	money	feeling	people	doing
SMOKING	crack	water	went	friend	body	
	smoke	soda	thought	time	eat	
	smoking	baking	house	weed	weight	
	pipe	freebase	car	smoking	eating	
	hit	spoon	shit	high	food	
	rock	rock	home	says	help	
SNORTING	nose	dry	smell	feel	nose	coke
	window	filter	card	coke	pain	line
	water	plate	bathroom	heart	damage	lines
	nasal	paper	coke	felt	blood	nose
	spray	powder	white	feeling	cocaine	small
	mouth	fine	bag	time	problem	cut
	MDMA (ECSTASY)					
GENERAL	time	serotonin	music	mdma	drug	pills
	really	mdma	rolling	experience	drugs	mdma
	first	effects	rave	time	mdma	pill
	feel	dopamine	people	people	people	test
	friend	brain	great	experiences	effects	ecstasy
	doesnt	receptors	mp3	feeling	depression	pure
	LSD (ACID)					
GENERAL	time	lsd	music	trip	experience	lsd
	acid	effects	tripping	experience	people	blotter
	friends	mescaline	movie	tripping	mind	blotters
	trip	psychedelic	love	time	think	dose
	friend	receptors	listening	first	lsd	taste
	felt	visual	watch	trips	way	dox
	OPIOIDS					
GENERAL	dont	pods	heroin	feeling	depression	dose
	know	tea	life	feel	drug	tolerance
	people	opium	years	time	drugs	opiates
	think	poppy	time	felt	treatment	opiate
	really	seeds	day	really	patients	high
	youre	pod	money	high	effects	doses
INJECTION	needle	water	dope	minutes		codeine
	vein	filter	time	later		pills
	veins	solution	shit	added		apap
	injecting	liquid	bag	seconds		liver
	blood	powder	know	hours		cwe
	hit	heat	going	10		acetaminophen

Table 4: Example output from a sample of pertinent delivery methods from five drug types. Darkened boxes indicate sparse tuples in which $b < 0.2$.