Mining the Web for Emerging Trends in Substance Abuse

OVERVIEW

Over the past year, NORC at the University of Chicago has been actively exploring the possibility of mining previously untapped online data sources to generate unique social and behavioral insights relevant to the field of substance abuse research. An initial investigation identified several online community forums dedicated to the discussion of relevant topics related to the use of illegal narcotics as well as the abuse of legal drugs, where users with registered accounts use the forums as anonymous venues to ask and answer questions about particular substances. Conversation topics range from recreational enjoyment to health and legal concerns, and many active users discuss multiple substances extensively in different subsections of such forums, providing social scientists with the potential to use these data for insight into drug-related information-sharing processes, recreational drug habits, and contextual information such as risk factors, comorbidities, and the distribution of various polysubstance combinations. In order to explore the analytic potential of this freely accessible data, we downloaded samples from two comparable forums as preliminary case studies and developed custom scripts to parse out useful data from every post collected, including the text of the post itself, the registered user that produced the post, and the particular section of the forum in which it was posted. Using these variables, we then conducted a variety of exploratory analyses in an attempt to assess the utility of various analytic methods, explore trends, and identify patterns related to drug use that may warrant further investigation. Encouragingly, the results of our network, natural language, and time-series analyses reveal a compelling variety of such findings, and suggest that further use of these analytic methods may allow researchers to more easily and rapidly detect and address emerging trends in narcotics use that might otherwise quietly develop into broader public health dangers.

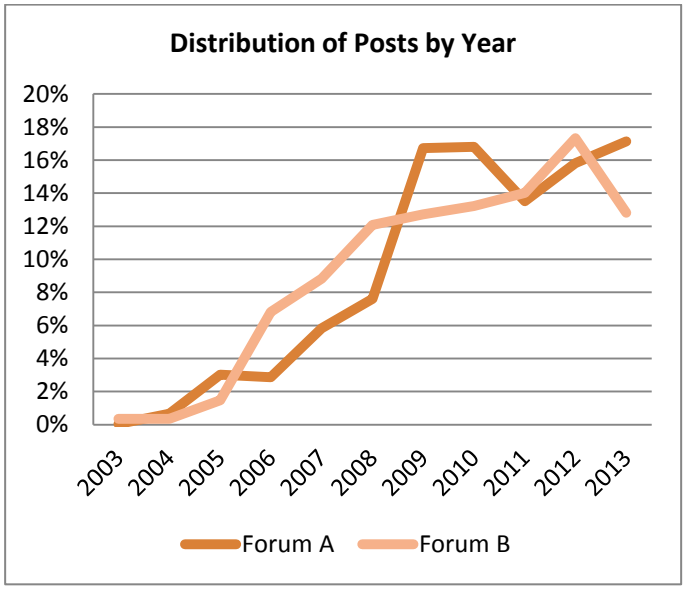
DESCRIBING THE DATA

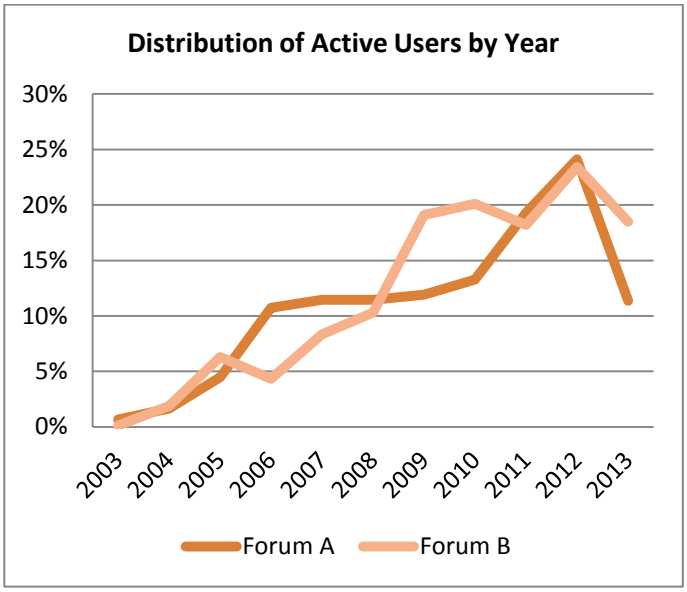
Using custom scraping and parsing scripts, over 100,000 posts were downloaded and parsed from the two case study forums to ultimately produce a dataset that captures years of online drug-related conversations, with topics ranging from opiate addiction and prescription pharmaceutical abuse to marijuana cultivation and the use of newly developed research chemicals. Our data on the two forums selected, referred to here as Forum A and Forum B[1], consist of 48,613 and 72,555 posts, respectively, with posts from each forum collectively produced by thousands of active users.

	Forum A	Forum B
Posts	48545	72336
Users	9071	3937
Avg. Threads per User	3.36	8.17
Avg. Posts per User	5.35	18.38

Figure 1 – basic descriptive statistics

[1] The names of these forums have been omitted for privacy protection purposes





Figures 2, 3 – distribution of posts and active users over time

While these basic numbers reveal nothing about the specific contents of each forum, they can serve as a useful starting point in understanding some of the core differences between the two case study communities. For example, while Forum A has fewer total posts than Forum B, its total user base is over twice that of the latter, resulting in a lower overall rate of user engagement. This suggests that Forum B is a more focused, exclusive community with more engaged users, while Forum A appears to be comprised of a larger pool of active participants that generate comparatively less content per-capita. While these differences alone may lend themselves to compelling lines of inquiry, for now they can simply demonstrate that each forum represents a unique online interest group, but that when taken together, they constitute a broad snapshot of two vibrant and active online communities where valuable and explicit information may exist regarding the use of illegal narcotics.

TRENDS IN TOPICS OVER TIME

As a first attempt to identify broad trends and better understand how dialogue across these data sources has changed in recent years, we began our analysis by examining changes in the volume of posts submitted to each forum over time, with posts categorized into broad topical areas that are already embedded into the structure of the websites themselves.

Volume of posts by category

Both forums are divided into a variety of sections or “subforums,” each dedicated to discussions pertinent to a specific drug-related topic. While these structures are intended to serve as a convenience for active forum users, they can also be used to categorize posts when conducting analysis. For example, when posts from each forum are divided into their topical categories and examined by volume over time, trends in the relative popularity of different topics can be made readily apparent. In examining the composition of Forum A since 2006, we find that overall activity has increased in recent years, and that the two most popular sections of the website are related to opiates and drug addiction, with a particularly substantial increase in the number of posts related to the latter over the past year.

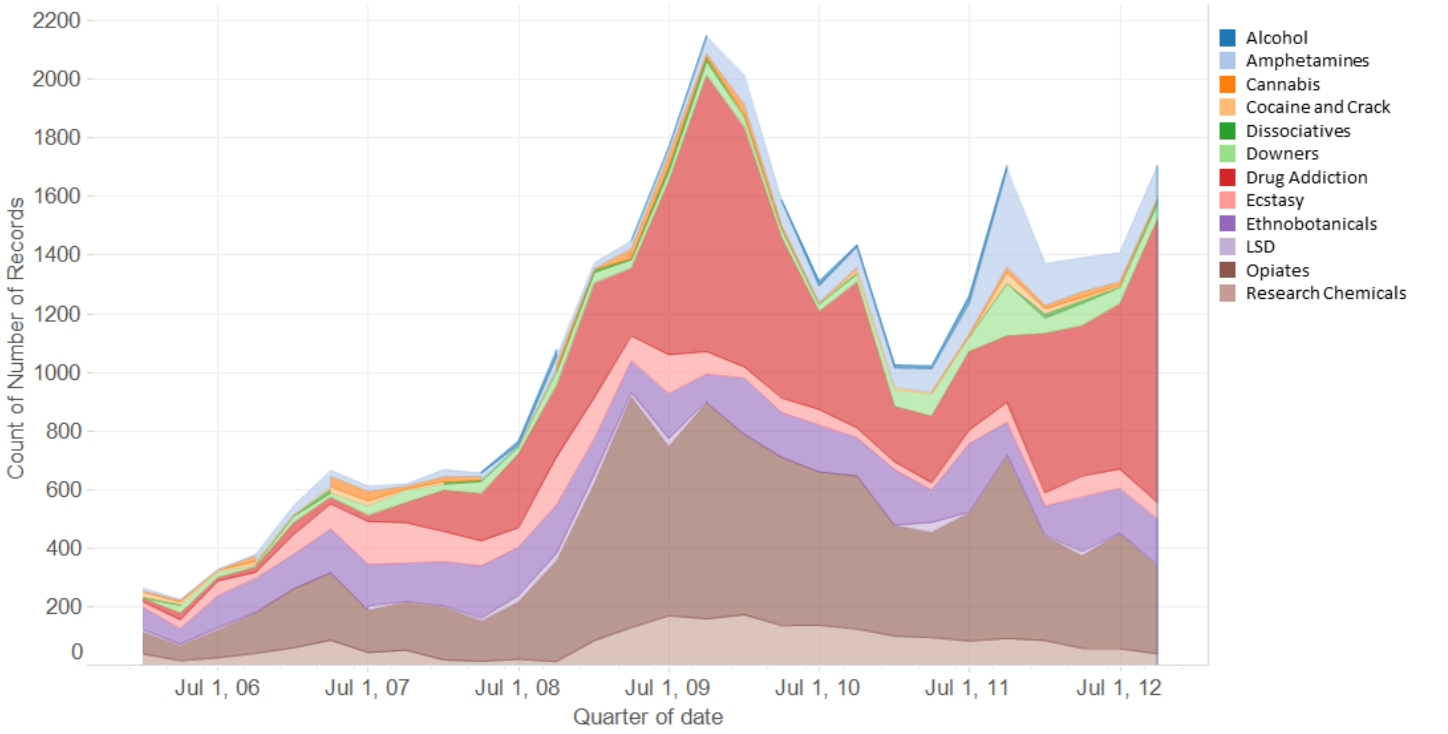


Figure 4a - total posting volume over time by section, Forum A

On the other hand, Forum B has enjoyed more consistent use over the past few years, and has only recently demonstrated an increase in overall posting volume. Curiously, previous spikes in activity appear to be periodic, a trend that may warrant further investigation. In contrast to Forum A, the section of Forum B related to discussion of Research Chemicals has begun to comprise a greater share of overall posts on the website since 2009, and its popularity appears to be rivaled only by the subforum related to Ecstasy.

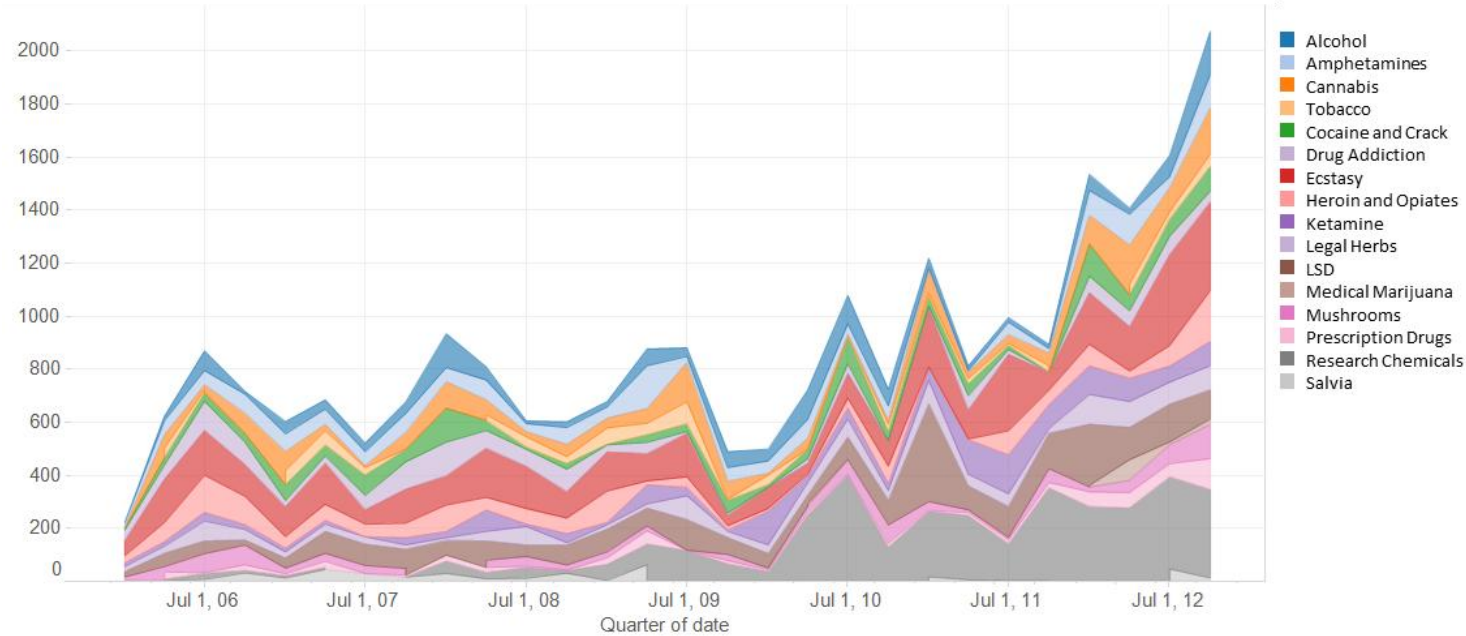


Figure 4b – total posting volume over time by section, Forum B

Composition of posts by category

We can also visualize these recent changes in posting composition in greater detail and conduct a more nuanced examination of forum dynamics by controlling for the overall volume of posts in each forum. By visualizing Forum A as an environment of competing topical

sections over time and disregarding the total volume of posts in any given year, the recent explosion in the relative share of posts submitted to the Forum A section on Drug Addiction becomes far more apparent.

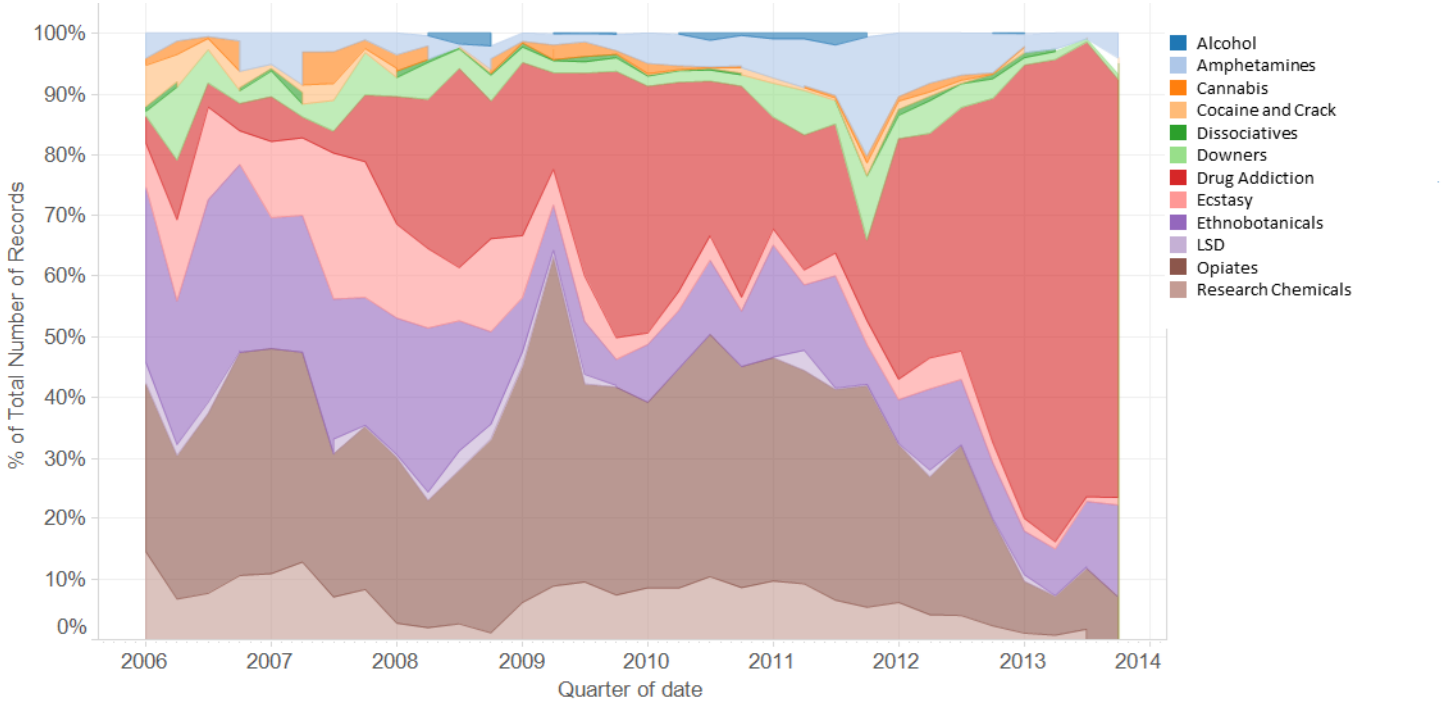


Figure 5a – percent of total posts by section over time, Forum A

Likewise, the long-term expansion of the Research Chemicals section on Forum B can be more easily compared to the forum's consistently popular section on Ecstasy in this manner, and a closer look also reveals a very recent emerging increase of posts related to Ketamine that appears to be at least partially responsible for the retracting popularity of Research Chemicals discussions.

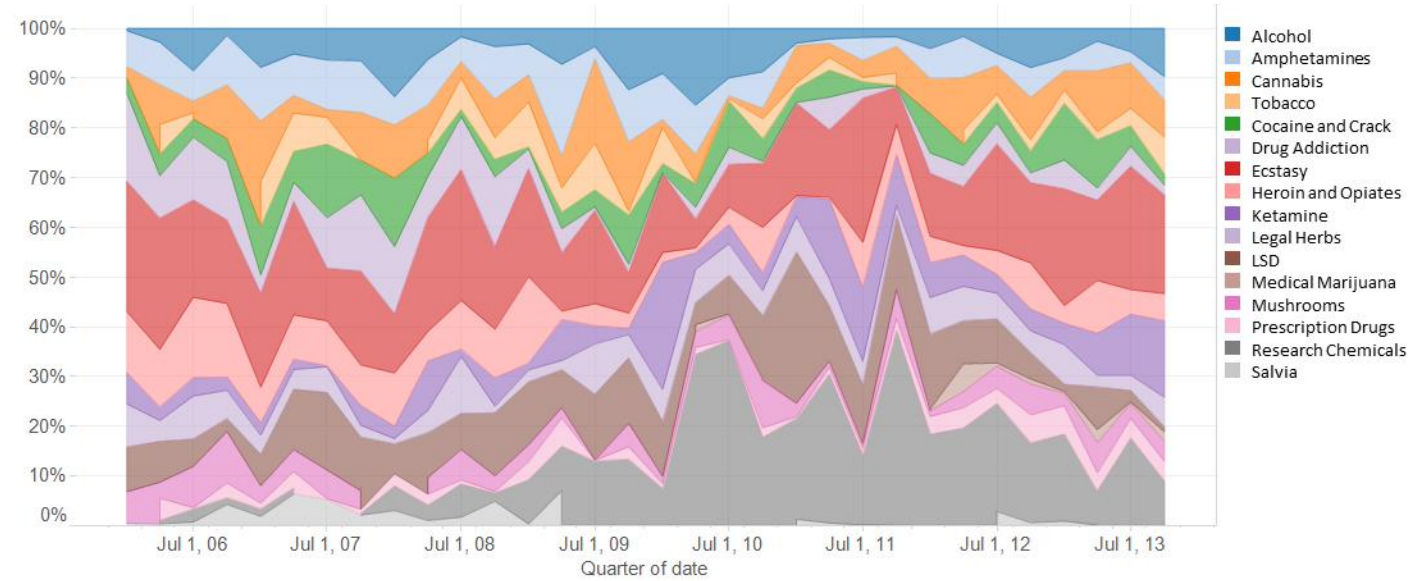


Figure 5b - percent of total posts by section over time, Forum B

Rapid changes in the annual composition of these forums can also be visually identified by using heat maps to highlight large shifts in the relative volume of posts in each section. Since 2008, Drug Addiction has clearly been the most popular section on Forum A, but it is also clear from this diagram that the dramatic increase in Drug Addiction posts in 2013 is uncharacteristically large even given its history, rising from 44% of forum volume in 2012 to 76% of total posts so far this year. This increase has also corresponded with a reduction in the share of Forum A posts in its second most popular section, Opiates, which has so far decreased from 23% to 8% of the

forum's overall posting volume from 2012 to 2013. Other notable spikes include brief increases in Amphetamine and Downers discussions in 2011, and a long-term gradual decline in the Ecstasy subforum since 2007.

	2006	2007	2008	2009	2010	2011	2012	2013
Alcohol			1.26%	0.13%	0.39%	1.12%	0.02%	0.05%
Amphetamines	1.76%	3.08%	1.93%	1.90%	4.94%	11.68%	7.97%	2.40%
Cannabis	2.35%	3.16%	1.55%	1.02%	0.79%	0.54%	0.78%	0.09%
Cocaine and Crack	3.69%	1.36%	0.38%	0.39%	0.33%	1.08%	0.63%	0.49%
Dissociatives	0.50%	0.62%	0.44%	0.56%	0.19%	0.04%	0.61%	0.69%
Downers	4.61%	4.56%	4.17%	2.15%	1.81%	7.21%	4.05%	1.18%
Drug Addiction	5.37%	6.28%	25.18%	32.01%	34.64%	18.17%	44.09%	75.97%
Ecstasy	10.07%	17.99%	13.30%	6.46%	2.74%	3.33%	4.07%	1.58%
Ethnobotanicals	29.70%	22.09%	20.73%	7.05%	10.01%	12.84%	10.14%	8.33%
LSD	1.68%	0.66%	1.96%	1.19%		0.84%	0.36%	0.56%
Opiates	30.37%	30.43%	27.11%	39.14%	35.23%	35.84%	23.25%	7.75%
Research Chemicals	9.90%	9.77%	1.99%	7.99%	8.92%	7.31%	4.02%	0.91%

Figure 6a - percent of total posts by section over time, Forum A

When examining Forum B in the same manner, Ecstasy and Research Chemicals emerge as the two dominant subforums, but the website also appears to have experienced a minor shift in composition towards a greater share of posts related to LSD in 2011, as well as an emerging resurgence of posts related to Cocaine and Crack in 2013.

	2006	2007	2008	2009	2010	2011	2012	2013
Alcohol	4.21%	6.42%	7.14%	6.04%	10.81%	2.20%	4.96%	5.12%
Amphetamines	7.97%	9.85%	7.48%	9.47%	6.93%	1.94%	5.64%	3.68%
Cannabis	6.27%	6.82%	7.61%	10.63%	2.32%	5.01%	7.92%	7.37%
Tobacco	1.98%	5.49%	2.89%	6.52%	2.42%	1.64%	1.71%	2.91%
Cocaine and Crack	4.09%	9.08%	6.66%	3.77%	5.80%	2.43%	5.13%	8.96%
Drug Addiction	11.76%	8.23%	11.56%	2.85%	1.92%	1.92%	3.49%	4.23%
Ecstasy	21.13%	18.28%	18.90%	15.60%	10.18%	17.67%	15.73%	21.36%
Heroin and Opiates	14.24%	8.31%	9.04%	6.41%	3.98%	3.86%	5.70%	5.94%
Ketamine	3.38%	2.30%	4.96%	4.08%	7.26%	10.64%	5.26%	7.71%
Legal Herbs	6.60%	3.75%	4.18%	4.94%	6.10%	4.99%	5.78%	5.46%
LSD	5.70%	10.77%	8.46%	11.42%	8.69%	18.18%	9.17%	3.50%
Medical Marijuana		0.32%			0.53%		1.74%	1.14%
Mushrooms	7.43%	4.28%	2.82%	2.23%	5.04%	2.66%	3.99%	3.71%
Prescription Drugs	2.10%	1.49%	1.63%	3.02%	1.03%	1.28%	4.20%	4.73%
Research Chemicals	1.53%	0.77%	4.11%	10.87%	26.97%	25.04%	18.72%	13.71%
Salvia	1.61%	3.83%	2.55%	2.16%		0.54%	0.85%	0.48%

Figure 6b - percent of total posts by section over time, Forum B

User engagement by category

Another means by which we can measure user engagement across these sections over time is to calculate the time spent between posts for each active forum user and then determine the average time delay across all posts in each subforum. In effect, this produces a measure that captures the relative posting frequency of users participating in discussions in each section, and can provide us with an alternative method to identify unusual trends in forum composition that would otherwise remain unnoticed by an analysis of overall posting volume. The diagrams below chart the average number of seconds elapsed between posts for users in each forum section over time, revealing sections where user engagement has been particularly lively.

On Forum A, we can see a substantial decrease in time between posts in the Alcohol and LSD sections from 2011 to 2012. While these are two of the smallest topical subforums on the website, their unusually high rate of user engagement appears to suggest that

the users that post in these sections are some of Forum A's most active. In particular, despite the overall proportion of posts related to LSD dropping substantially in 2012, the posting frequency of LSD subforum users increased drastically over the same period, indicating that while other interest groups enjoy far greater volume of discussion on the website, users in the LSD section continue to grow more distinguished as the most active group on the Forum A message boards.

	2006	2007	2008	2009	2010	2011	2012	2013
Alcohol			2,697,842	3,807,687	2,879,527	5,593,414	153,000	497,880
Amphetamines	1,618,074	1,770,747	1,462,916	2,426,309	2,568,980	2,349,426	2,091,218	2,262,333
Cannabis	2,387,466	1,268,655	2,282,647	4,515,700	3,403,265	5,416,420	4,465,088	1,535,376
Cocaine and Crack	2,475,363	1,056,467	2,723,280	2,274,256	1,676,697	2,396,516	4,740,130	2,183,076
Dissociatives	2,715,590	2,321,444	765,724	2,641,955	3,427,545	3,112,530	1,527,420	363,313
Downers	1,417,261	3,343,285	1,896,299	2,159,068	3,553,467	4,132,361	4,434,306	4,707,194
Drug Addiction	2,577,312	2,954,551	2,155,056	1,178,171	1,349,328	2,376,236	2,086,268	1,405,256
Ecstasy	2,120,522	2,183,837	2,562,772	1,609,354	2,121,515	3,816,892	1,677,905	2,144,359
Ethnobotanicals	2,326,621	1,881,088	2,216,855	2,545,307	2,670,820	2,138,850	3,594,752	1,461,160
LSD	4,837,167	1,267,035	2,608,353	3,607,062		3,281,716	497,054	269,400
Opiates	1,512,864	2,139,021	2,104,578	2,002,285	2,249,555	2,944,672	2,804,814	2,396,530
Research Chemicals	2,932,146	2,791,793	5,143,416	2,344,115	2,582,423	4,427,811	3,970,514	3,403,164

Figure 7a – average time delay between user posts by section over time, Forum A

Posting activity across Forum B's subsections appears to be more consistent, particularly in recent years. Notable exceptions include a substantial slowdown in posting frequency in 2011 among users posting in the Ketamine section, as well as a recent and drastic drop in the average posting frequency of users participating in conversations in the Tobacco subforum in 2013.

	2006	2007	2008	2009	2010	2011	2012	2013
Alcohol	299,248	738,282	682,435	606,066	361,582	745,697	505,023	61,046
Amphetamines	436,023	267,627	469,172	602,553	416,849	588,885	416,201	301,770
Cannabis	895,700	549,398	827,566	214,948	353,367	242,405	198,517	221,157
Tobacco	1,207,658	831,454	256,271	287,932	439,142	514,842	99,084	1,260,454
Cocaine and Crack	748,052	331,945	455,351	179,025	475,080	208,763	315,741	304,282
Drug Addiction	175,876	813,659	568,177	522,966	1,373,847	579,133	581,437	121,510
Ecstasy	460,723	375,175	455,629	378,193	424,444	594,492	484,084	392,433
Heroin and Opiates	145,808	564,300	308,484	222,740	74,246	171,520	103,349	204,732
Ketamine	459,463	1,270,051	320,404	136,512	168,193	1,396,242	483,377	190,251
Legal Herbs	970,222	859,126	417,812	194,252	288,763	238,636	156,651	193,275
LSD	243,916	743,229	334,651	288,604	503,076	227,026	410,660	372,166
Medical Marijuana		241,568			95,584		109,876	146,407
Mushrooms	251,040	259,484	1,048,557	1,350,613	69,877	202,094	333,806	407,540
Prescription Drugs	232,555	602,810	193,911	421,850	60,248	288,541	349,429	256,278
Research Chemicals	380,274	231,973	1,109,529	255,722	342,625	359,797	524,964	469,593
Salvia	603,085	678,641	765,688	600,687		69,254	62,124	253,280

Figure 7b – average time delay between user posts by section over time, Forum B

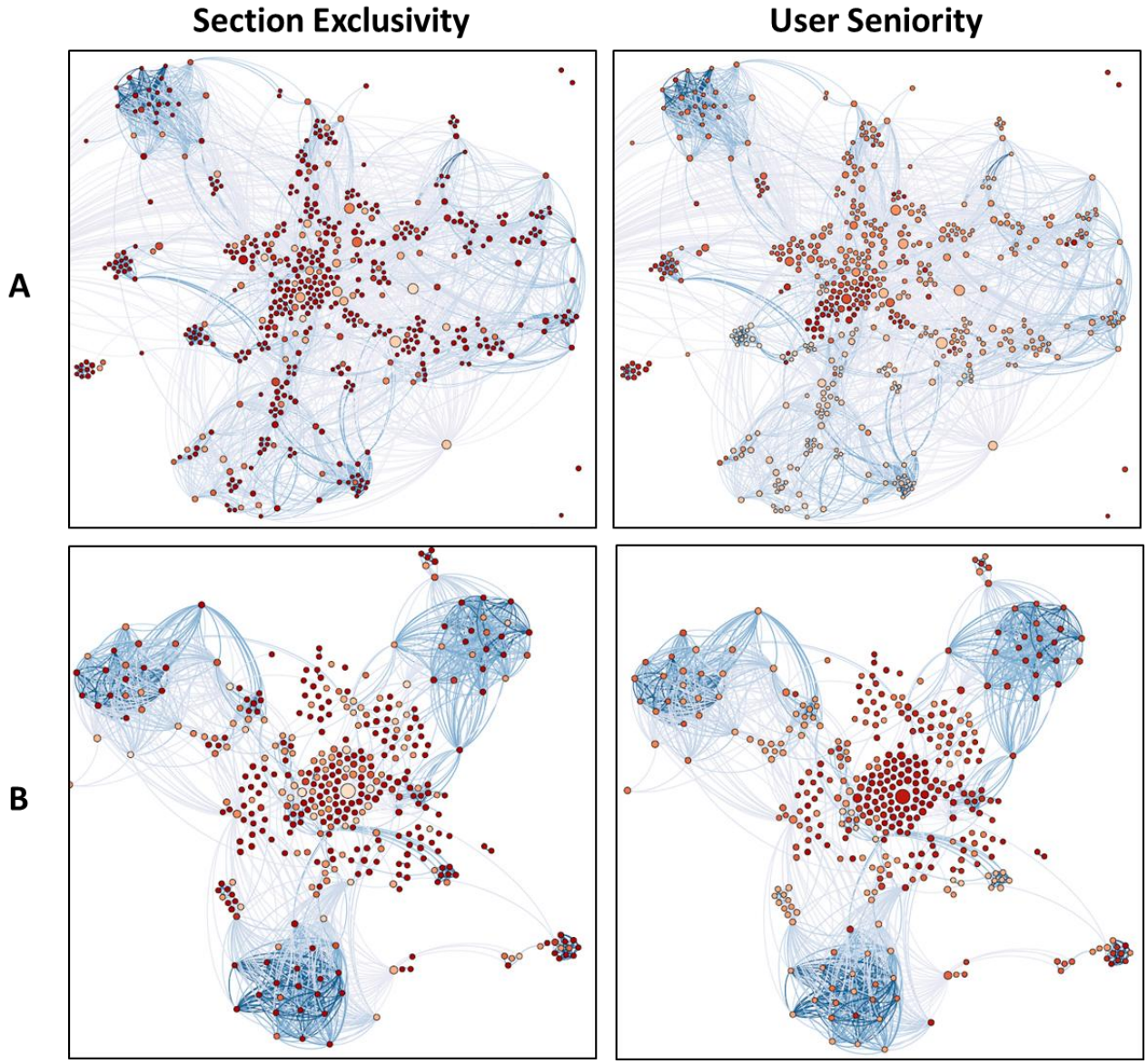
TOPICAL INTEREST GROUP NETWORKS

More nuanced information about user interactions in these subsections can be gained by transforming these datasets into social network diagrams, transitioning the unit of analysis from posts to the users themselves. With each forum composed of sequential groups of user-generated posts called "threads" that are themselves organized into sections for specific topics, posts can be aggregated into a structure suitable for analysis as a social network by calculating each user's distribution of posts across every section to determine the subsection in which they have most often posted, or what could be considered their "favorite section." Using this information, an analysis can then be conducted on the subset of users that post more frequently in each section relative to the other sections on Forum A and Forum B, and users within each section subset can be examined in relation to one another based on measures calculated from their posting activity. While there are many measures that could be used to do so, a simple method of linking users together is to determine whether two users have both participated in a mutual thread before, and if so, to calculate the proportion

of total threads between the two that they have shared. In effect, this simple measure can provide us with a rough estimate of the degree to which two users participate in the same discussions as one another.

By linking users together in this manner to generate network diagrams, we can better understand how drug-related information is exchanged on these websites within certain topical scopes. In the diagrams below, each point or "node" represents a user, and links between users indicate that the two have both posted in at least one of the same discussions. Users appear closer together if they have posted in the same threads for a greater proportion of their total posting activity, and the size of each user node is based on the number of total posts that they have made across the whole forum website. As indicated by the captions, node colors represent either the proportion of the user's total posts that were made in the section (indicating the degree to which the section is their "favorite" relative to other sections of the forum), or the time of the user's first post (indicating how long they have been active on the website.)

Ecstasy subforum networks



Figures 8a, 8b, 8c, 8d – network of users that prefer posting in the Ecstasy sections of Forum A (top) and Forum B (bottom); coloring (left) indicates the percent of a user's total posts that were made in the section, with darker red indicating higher posting exclusivity to this section; coloring (right) indicates time of a user's first post, with darker red indicating that the user joined the forum more recently

The Ecstasy subforum on Forum B appears to consist of several devout clusters of users, with their dark coloring indicating that these groups post almost exclusively in the Ecstasy section. These clusters of users appear to be connected to one another by other users that participate more actively in other sections of Forum B (as indicated by their lighter coloring), serving as "bridges" between the

Esctasy-only users. One such bridge stands out in particular with a much lighter exclusivity shading and larger size than the others, indicating that this user participates actively in the other Forum B subforums and the user has generated a much higher total volume of posts than other Ecstasy section users, despite being relatively new to the forums; it is likely that this node represents an emerging leader or new administrator.

On the other hand, users in Forum A's Ecstasy section that demonstrate higher and more diverse posting activity across other sections appear to have a high degree of centrality, and do not act as "bridges" between distinct groups of users to the extent demonstrated by those in Forum B's Ecstasy subforum. In general, user participation in this section on Forum A appears much more diffused across the whole section, although Forum B's newer Ecstasy users are found mostly in the main clusters and "hubs" of activity, in contrast to Forum A's Ecstasy section, in which they tend to be spread out on the periphery.

Research Chemicals subforum networks

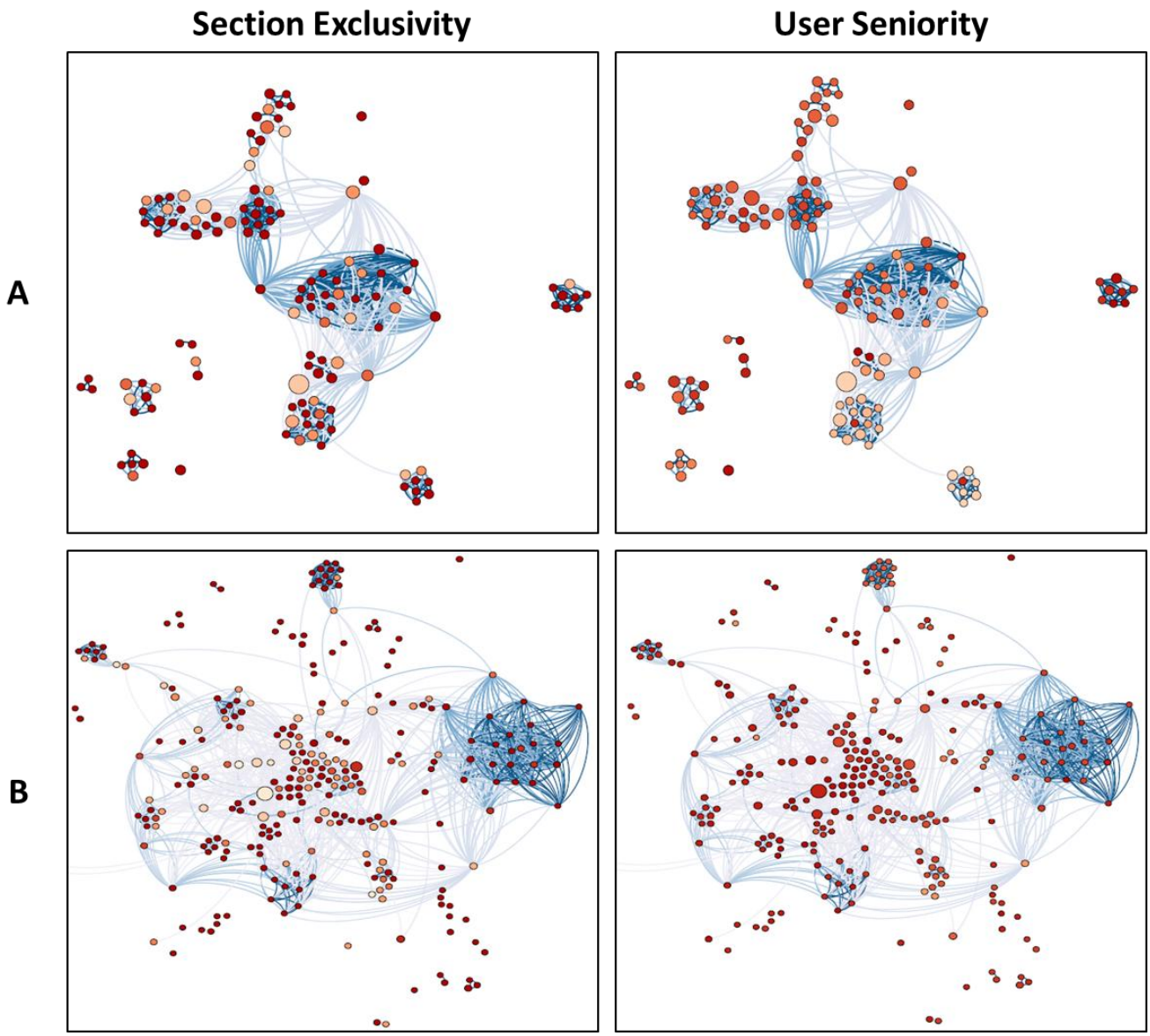


Figure 9a, 9b, 9c, 9d – network of users that prefer posting in the Research Chemicals sections of Forum A (top) and Forum B (bottom); coloring (left) indicates the percent of a user's total posts that were made in the section, with darker red indicating higher posting exclusivity to this section; coloring (right) indicates time of a user's first post, with darker red indicating that the user joined the forum more recently

In comparison, the sections dedicated to discussions of Research Chemicals in both forums appear to be small and highly clustered, centered in particular around "devout" posters that appear to be grouped by their seniority. As seen previously, a few select members serve as "bridges," engaging in conversations with groups of users that appear to otherwise keep to themselves. Also notable is the

presence of a dominant “leader” node with a high degree of centrality; while this was previously seen in the Forum B Ecstasy section network, this type of node is present here in both forums. This user appears to not only engage with other sections of the forums more than his peers, but has also been active on the forums at least as long (in the case of Forum B) or far longer (in the case of Forum A) than many other users, suggesting that this user plays a key role in the evolution of discussions relating to Research Chemicals.

Despite their simplicity, these preliminary network analyses clearly illustrate the potential of this methodology in developing new understanding of how users engage with one another around different drug-related topics in online forum environments. The nature of interactions between users appears to vary substantially across different groups of users in different topical subsections, and further analysis could allow us to easily identify central leaders in each of these interest groups. Such information could then be studied in conjunction with users’ posts through natural language processing, allowing researchers to focus on key network nodes and track the dissemination of emerging topics and trends with greater efficacy.

NATURAL LANGUAGE ANALYSIS

However, analysis of these topical subforums need not be limited to aggregate posting statistics and particular users. Indeed, a broad examination of the language of users on these forums reveals much about the nature of the emerging trends identified earlier. To better explore the specific content embedded in these collections, we cleaned and de-duplicated posts for each forum (removing punctuation, expanding contractions, correcting common spelling errors, etc.), and created a dictionary of hundreds of keywords related to notable drug-related topics, including slang terms for popular narcotics, names of chemical compounds, terms related to health issues, keywords related to specific types of drug use methods, adjectives commonly used to describe drug use experiences, and so forth. We then examined the prevalence of these keywords over time in order to capture how discussions on these two forums have been changing recently, including the use of three-dimensional correspondence analyses, which cluster words around particular years based on the frequencies and case occurrences of the selected keywords.

Changes in dialogue over time

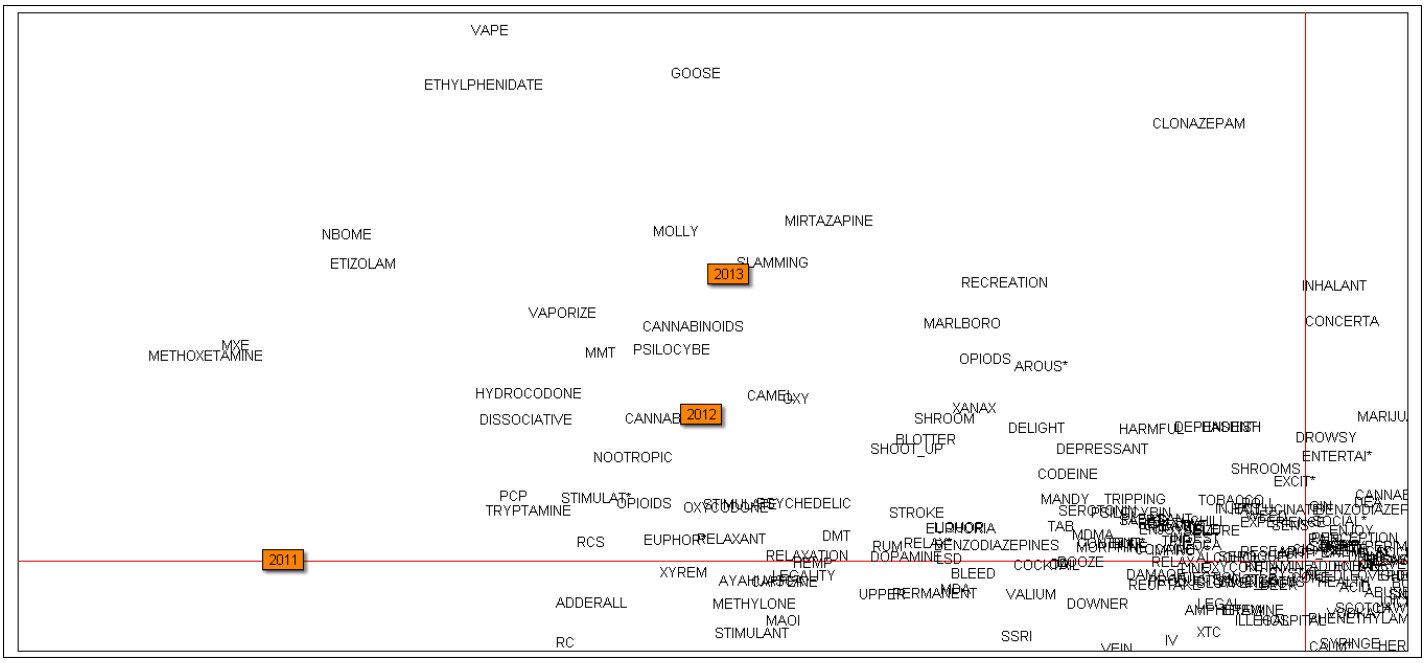


Figure 10a – drug-related keywords most distinctive of particular years, Forum B

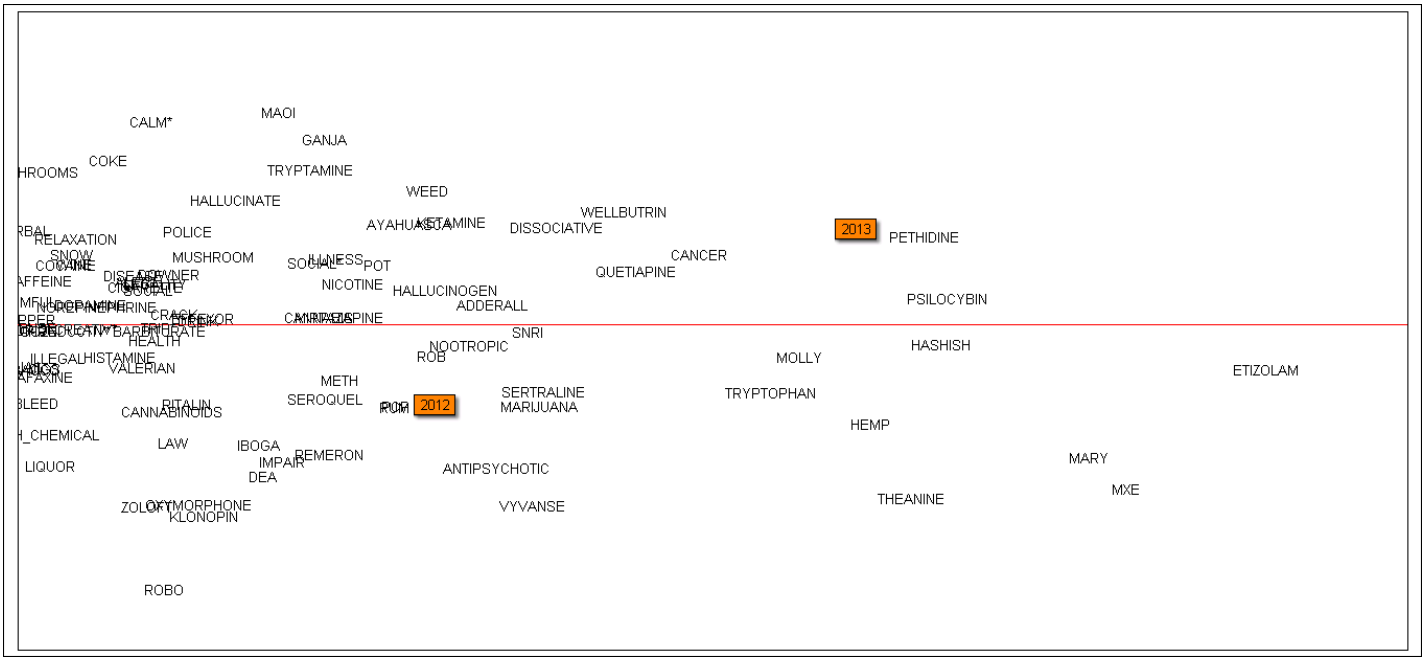


Figure 10b– drug-related keywords most distinctive of particular years, Forum A

As can be seen, some unique new words have been emerging in Forum A in 2012 and 2013 that are unlike those used in past conversations, including “pethidine,” “etizolam,” “mxe,” “theanne,” and “tryptophan.” Similarly, emerging words for Forum B include “ethylphenidate,” “etizolam,” “nbome,” “mirtazapine,” “slamming,” “goose,” “mxe,” and “methoxetamine.” Given these compelling findings, we believe that continued analysis of this nature may serve as a very effective method of identifying new slang terms and new drug trends early on. For example, the earliest use of the term “molly” appeared on Forum B in 2006 and began to appear in multiple posts as early as 2008. In early 2010, the word began appearing on both Forum B and Forum A, and would have likely been detectable soon after through the use of a natural language methods such as those demonstrated above.

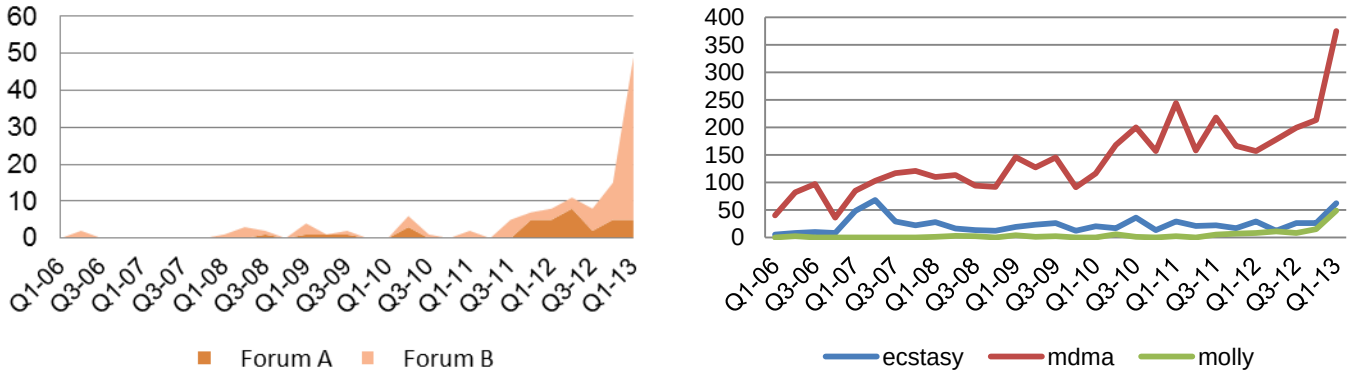


Figure 11a, 11b – occurrences of the term “molly” by forum (left) and total MDMA-related keyword frequencies for both forums (right)

Likewise, terms related to opiates became extremely prevalent on both forums in late 2008, but were emerging as early as Q1 2007.

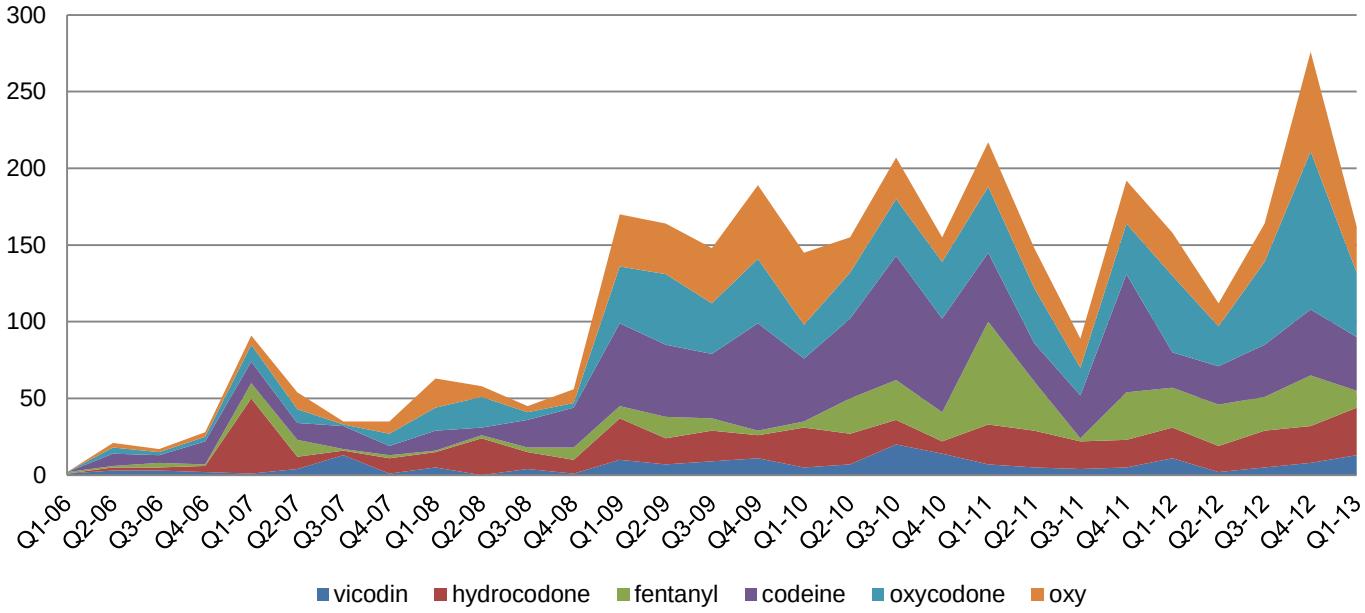


Figure 12 – occurrences of opiate-related terms for both forums

Identifying topical clusters

Particular words of interest can also be scrutinized further in order to gain insight into the contexts in which they are used, captured by other terms with which they tend to be found in close proximity. For example, ordered agglomeration can be used to explore the words that are most closely related to the word "caffeine" based on the degree to which they are found together in the same post.

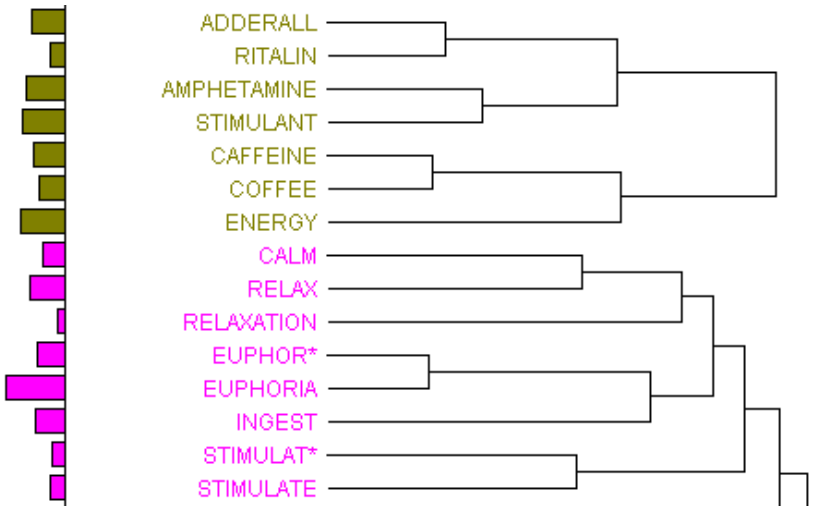


Figure 13a – dendogram clustering of the word "caffeine," Forum A

In Forum A, "caffeine" is – unsurprisingly - most commonly found in conjunction with the word "coffee." The next strongest link to another word is found with "energy," and beyond that pairing, the "caffeine/coffee/energy" cluster is most often found in proximity to another cluster characterized by the stimulant-related words "Adderall," "Ritalin," "amphetamine," and "stimulant." In Forum A, then, it appears that the use of caffeine tends to be linked to use of prescription "study drugs." However, the association is different in posts from Forum B.

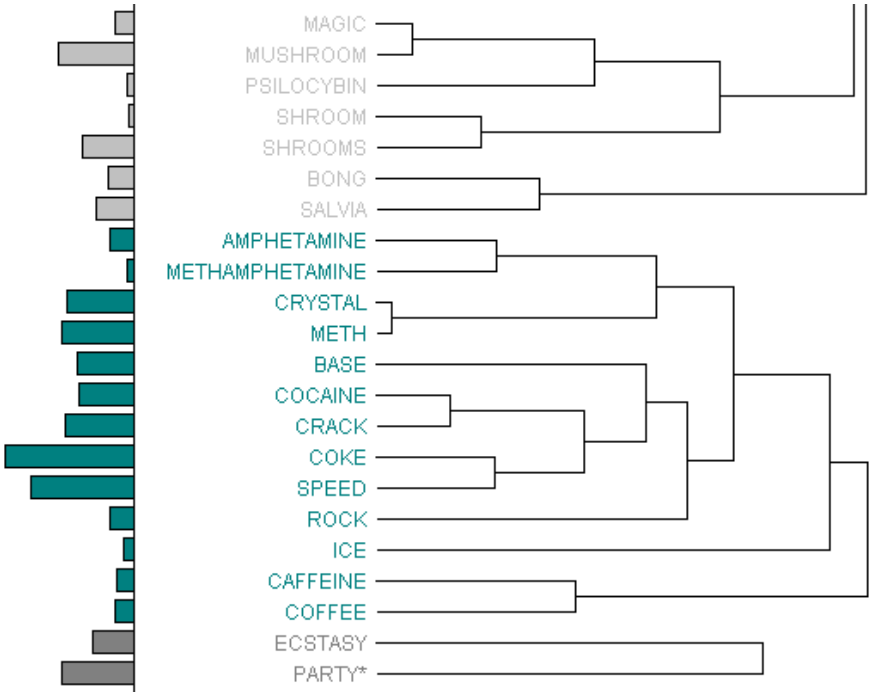


Figure 13b – dendogram clustering of the word "caffeine," Forum B

As with Forum A posts, mentions of "caffeine" in Forum B are most commonly associated with "coffee," but the next largest cluster includes mentions of "cocaine" and "speed" rather than prescription pharmaceuticals. It would appear, then, that use of caffeine on Forum B is discussed in terms of recreation rather than productivity, or perhaps in regards to its ability to keep users awake rather than improve their concentration.

Similar analysis of other drug-related terms could very well reveal much about the relation of particular substances to one another in terms of the natural dialogue generated by users of online drug forums. However, this method of clustering requires the prior selection of particular keywords, and it may also be useful to examine the forums in terms of words that appear to be naturally distinctive in their native contexts. For example, particular sections of each forum can be used as categories and can be applied to a classification algorithm in order to determine the words most distinctive of each section. Below are terms that are found significantly more often in one section of Forum A or Forum B than in any other.

Forum A			
Downers	Addiction	Opiates	Research Chemicals
CLONAZEPAM	LIFE	TRAMADOL	METHYLONE
BENZO	HELP	HEROIN	MEPHEDRONE
ALPRAZOLAM	QUIT	OXYCODONE	COMPOUND
DIAZEPAM	TAPER	FOIL	CHEMICAL
XANAX	ADDICT	INJECT	VENDOR
BENZOS	RELAPSE	MORPHINE	STIMULANT
BENZODIAZEPINES	TURKEY	IV	DURATION
ANXIETY	SYMPTOM	VEIN	MIN
PRESCRIBE	QUITTING	SNORT	NOSE
VALIUM	GBL	SHOOT	HEART
MG	COLD	OXY	ORALLY
MEDICATION	HABIT	FENTANYL	ORAL
SLEEP	RELATIONSHIP	OXYCONTIN	SLIGHTLY
DOCTOR	HANG	SYRINGE	ROUTE
BRAND	REALISE	HYDROCODONE	
INSOMNIA	RECOVER	GEAR	
PANIC	ANYMORE	CRUSH	
COMBO	IM	OD	
NATURAL	HATE	PATCH	
MUSCLE	STOPPING	RUSH	
MANAGEMENT	SUB	SHOO	
CALM	BOYFRIEND	NALOXONE	
HRS	BUPRENORPHINE	NEEDLE	
WONT	WITHDRAWL	CAREFUL	
	BREAK	POD	
	BUPE	POPPY	
	SLOWLY		
	PARTNER		
	CONSTANTLY		
	FIX		
	ANSWER		
	DIDNT		

Forum B		
Addiction	Prescription Drugs	Research Chemicals
ADVISE	ONLINE	WTF
HARD	BENZOS	SIMILAR
GROUP	DOCTOR	AMT
GOD	TABLET	MXE
HOPE	DXM	BENZO
LIFE	ANXIETY	APB
HUG	LIVER	MEPH
SUPPORT	MG	BATCH
ADDICTION	VALIUM	VENDOR
SYMPTOM		NRG
WEEK		REPORT
BOUNCE		MEPHEDRONE
BACK		STUFF
METHADONE		HO
ADDICT		SLIGHTLY
SUBUTEX		DRONE
DAY		STIMULANT
GEAR		EUPHORIA
HELP		COMPARE
CARE		MILD
FAMILY		RC
FL		
STORY		
HABIT		
LUCK		
SITUATION		
CALL		
MANAGE		
PV		
FINALLY		
FORGET		

Figure 14 – words significantly associated with one section relative to others, Forum A (left) and Forum B (right)

By classifying words in this manner, certain associations become immediately apparent – some intuitive, and others more provocative. For example, “clonazepam” is understandably found disproportionately in the Forum A section dedicated to Downers, and “methylone” is significantly distinctive of the section on Research Chemicals. However, while these findings may be unsurprising, they can point researchers towards the keywords that best represent specific topical sections of each forums, and also inform a better understanding of how substances are discussed differently across separate online communities; “benzo” may be considered a topic for Downers on Forum A, but on Forum B it is considered appropriate for Research Chemicals. A closer look also reveals other more notable findings, including that the term “methadone” is present in Forum B’s section on Addiction significantly more often than in its section on Opiates, where one might expect it to be used most commonly, suggesting that addiction to methadone may be a particularly distinct topic on Forum B.

Given the example above, it might make sense to search for topics related to the keyword “benzo” without restricting the analysis to a particular subforum, and in a manner that would allow us to measure its distribution across the entire website in addition to simply identifying other terms with which it is commonly associated. Fortunately, for more insight into the broader themes and topics across these discussion boards, we can also make use of algorithms that do not require predetermined keywords or sections, allowing for the identification of the most distinctive natural word clusters without restricting the data in any way. For example, the following 25 topics were derived for each forum using Latent Dirichlet Allocation, which attempts to characterize posts in terms of a preset number of word clusters and then allocate these “topics” across the full collection of posts.

Topic	Mean	Median	Max	Min
buy stuff get make meth chemical small cut crystal way sell powder used quality easy put bottle form amount	3.92%	3.51%	90.12%	0.06%
thing like feel something make maybe way mind anything find really else cannot sometimes hard time think seem know	4.11%	3.64%	43.72%	0.05%
good know really get though pretty bit stuff like sure going much think nice well try anyway better quite	4.17%	3.64%	50.32%	0.05%
love crazy wink quote posted group bounce well good originally weee yeah sorry hug sound mate fl thats you	3.98%	3.57%	47.83%	0.04%
know guy girl want sex never friend really someone say think like relationship thing love person tell always talk	4.00%	3.57%	32.73%	0.06%
quot party even people name false cop area london scene rave club folk old often event also today place	4.01%	3.51%	97.43%	0.07%
drug state said year government police heroin rsquo news new cannabis policy war report country market control bbc city	3.99%	3.51%	80.06%	0.01%
mdma like acid trip experience dmt lsd never tried done time really pretty much bad sound though ever good	4.01%	3.57%	50.93%	0.04%
still went could started back felt like around friend feeling thought got time night bit hour came eye remember	3.94%	3.54%	38.54%	0.06%
people drug even life problem many age lot year others young kid still use family etc issue seen school	4.02%	3.57%	50.91%	0.01%
think would thing one like way look doe make might see idea come part dog actually know kind men	4.10%	3.64%	71.96%	0.04%
get take day help need drink alcohol heroin stop pain drinking not taking know month will much every opiate	3.96%	3.57%	49.63%	0.06%
laugh at lol man like head yeah look funny guy thought haha dude put face hand big said one	4.10%	3.64%	66.67%	0.03%
one pill year got ago two old back would party name well found around friend month mine round going	4.02%	3.57%	56.76%	0.06%
mushroom like yes good eat need make look best bit water nice food plant taste get little etc seed	3.91%	3.57%	72.47%	0.05%
time day first last night mg hour week half next took take line got gram every long weekend dose	3.96%	3.51%	75.09%	0.09%
dont shit mate fuck like lol thats cant fucking get got cos tho fucked coke know load ket yeah	3.91%	3.57%	63.71%	0.04%
weed smoke smoking crack used cocaine drug white marijuana hit smoked high price heroin black blue bag lsd red	3.93%	3.51%	71.34%	0.05%
effect use drug may test doctor cause research brain ecstasy substance user death term also case risk medical result	4.05%	3.57%	54.55%	0.08%
one world say new free believe made well show right would must way human point mean set yet word	3.96%	3.57%	40.00%	0.05%
bad like feel sleep really good effect taking much day take bit speed long comedown keep thing heart going	3.93%	3.57%	54.55%	0.10%
drug use illegal country user law class many chemical case also substance often legal risk s dealer service less	4.08%	3.51%	38.41%	0.02%
http post www forum read com thread site thanks amp anyone please watch partyvibe looking find info help link	3.85%	3.39%	86.87%	0.02%
would get lot much high stuff probably legal though think enough even actually people still getting better might although	4.03%	3.64%	38.97%	0.04%
would want work get said going know need someone see could say right getting people let back trying something	4.04%	3.64%	35.90%	0.06%

Figure 15a – topic model of 25 word clusters for all posts, Forum B

Topic	Mean	Median	Max	Min
drug effect brain may receptor cause patient disorder serotonin depression treatment increase anti ssri system dopamine level stu	3.88%	3.08%	80.78%	0.14%
drug addiction use life people addict problem using addicted person many think control self one need become issue change	4.23%	3.42%	41.84%	0.05%
day withdrawal week month mg dose methadone suboxone take long taking taper time detox last year every symptom mgs	4.32%	3.33%	37.50%	0.09%
mdma one long even experience would sex well way like much time least perhaps rather brain without might may	3.90%	3.39%	28.10%	0.13%
body anxiety also help sleep time problem take taking symptom still much side cause eat like even lot drink	3.93%	3.37%	32.71%	0.10%
day hour night feeling feel took last felt time still sleep got went first next back today started morning	4.22%	3.39%	37.44%	0.07%
get like know really want shit dont going got even right never thats thing bad though ever lol fuck	4.21%	3.57%	36.40%	0.01%
pain doctor get take would need help work med prescribed back doc medication said taking problem well know could	3.98%	3.33%	33.80%	0.04%
amp drug cocaine legal new people also name substance state product illegal available case area many street cost used	3.77%	3.21%	38.39%	0.10%
swim doe time would like tried good never first try feel get pill still experience could much friend found	3.96%	3.28%	28.95%	0.01%
swiy get help keep going good need clean best want hope feel stay luck done know getting really well	4.25%	3.51%	32.00%	0.04%
quot people say one would said even person make could call like someone mean man cannot saying word think	3.89%	3.33%	49.63%	0.09%
really thing like get feel good make much think lot better something try work maybe doe not will bad	4.21%	3.61%	27.11%	0.03%
post comment reputation forum thread read please drug question information site find member quote http answer www info com	4.06%	3.25%	56.76%	0.06%
like mind thought trip thing could something experience eye one time see around world head music real point heart	3.87%	3.33%	59.55%	0.11%
water powder tea leaf alcohol amount method extract taste gram use using make bottle capsule small opium liquid drink	3.83%	3.08%	75.09%	0.09%
also used one acid found plant vein use form many oil may juice made pure must seed make coffee	3.64%	3.13%	63.71%	0.14%
smoking smoke meth weed hit like get good one smoked way coke put line cigarette high little end pipe	3.99%	3.28%	51.52%	0.05%
friend year time life never got back started money month every could drinking old would family always went away	4.22%	3.45%	35.41%	0.08%
opiate heroin would one methadone high pill doe use much oxycodone morphine buprenorphine used oxy even hydrocodone tol	4.01%	3.33%	34.81%	0.05%
effect substance cannabis dmt chemical user experience alcohol lsd research similar report mephedrone however result less com	3.84%	3.23%	54.55%	0.16%
mg effect kratom tramadol dose take high tolerance experience taking dos cat euphoria dosage hour codeine taken tried like	4.14%	3.33%	38.41%	0.02%
drug rsquo marijuana said state ldquo treatment law rdquo year health court use medical study program new police hellip	3.58%	2.94%	86.87%	0.19%
know would test someone time doe want tell work could positive thing right drug see get way one people	4.00%	3.45%	38.97%	0.06%
would thanks second later know minute think added much like anyone could sure say thank good also one heard	4.04%	3.51%	35.90%	0.08%

Figure 15b– topic model of 25 word clusters for all posts, Forum A

While some of these topics contain words related to general use (e.g. “really thing like get feel good…”), others capture more focused lines of conversations, such as healthcare (“pain doctor get take would need help work med prescribed…”) and opiate use (“opiate heroin would one methadone high pill doe use much oxycodone morphene…”) Once generated, the prevalence of these topics can be

tracked over time, along with the degree to which they continue to characterize new content – or fail to do so. Topic models can also be fitted to subsets of data, such as recent posts from each forum in a specific year.

Language distinctive of new forum users

More nuanced analysis can also reveal additional information about emerging linguistic trends across the forums. Keywords commonly used by newly registered users, for example, might be one particular subject of inquiry that could yield useful insight into topics popular among newcomers relative to veterans of drug-related online discussion forums. Using a Chi-squared test, we can identify the top 20 keywords that are significantly associated with new users' first posts in each forum, representing a snapshot of the interests that first bring new users to these websites:

Top 20 Most Significant Words for New Forum Users	
Forum A	Forum B
SMOKE	MDMA
WEED	EXPERIENCE
MEPHEDRONE	ENJOY
SNORT	LEGAL
ROLL	SENS*
MDMA	SPEED
CRYSTAL	SAFE
COKE	PERC
EUPHORIA	TRIP*
OXYCONTIN	CRACK
METH	HANGOVER
WITHDRAWL	HIT
XTC	METH
MOLLY	EAT
EXPERIENCE	INGEST
HIT	EXCITE
SWALLOW	SNORT
XANAX	SHROOMS
FUN	DEXTROMETHORPHAN
LIGHT	IV

Figure 16 – top 20 keywords most significantly associated with first posts by new users on each forum

A number of substances emerge as particularly distinctive of newcomers to Forum A and Forum B, suggesting that drugs such as MDMA, mushrooms, methamphetamines, oxycontin, and Xanax may serve as gateways, curiosity about which may draw individuals to these online communities. These trends can also be analyzed over time using the same three-dimensional correspondence analysis method demonstrated above. In the diagrams below, the ones and zeros preceding the year on each label indicate cases that represent the first post from new users and subsequent posts from existing users, respectively.

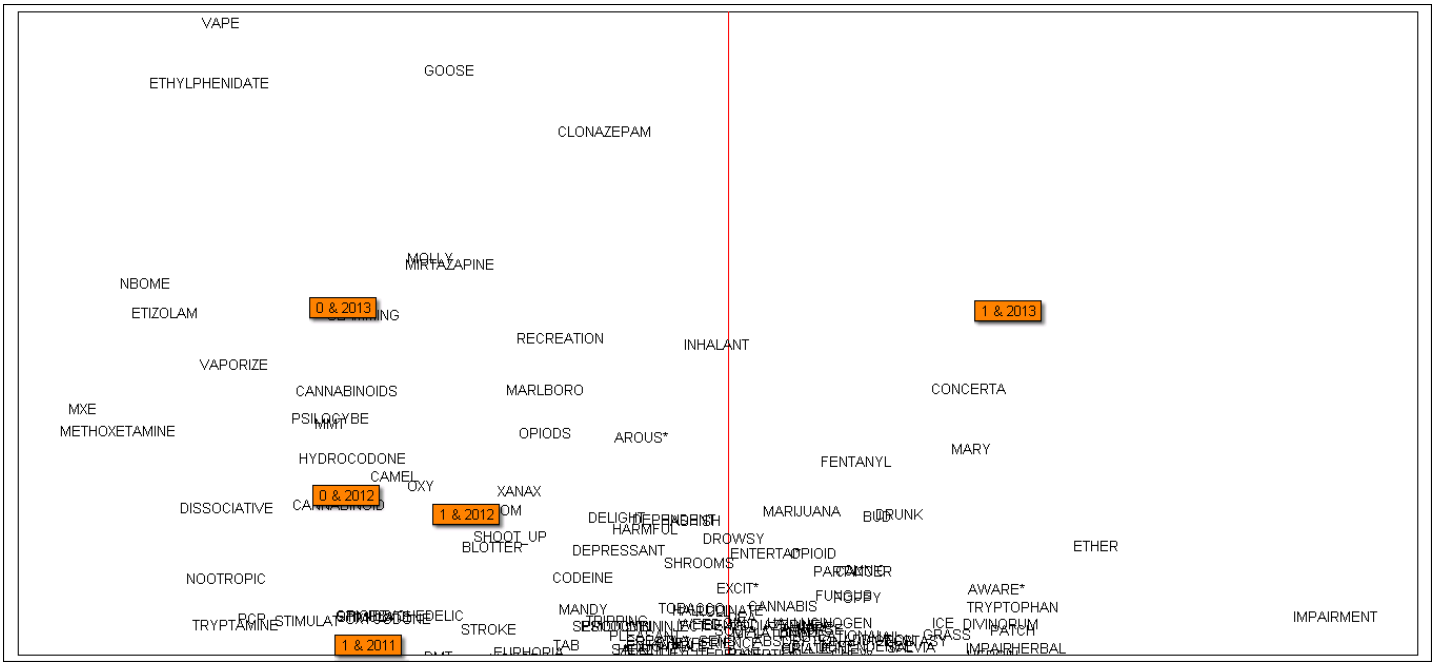


Figure 17a– drug-related keywords most distinctive of posts from new and existing users in particular years, Forum B

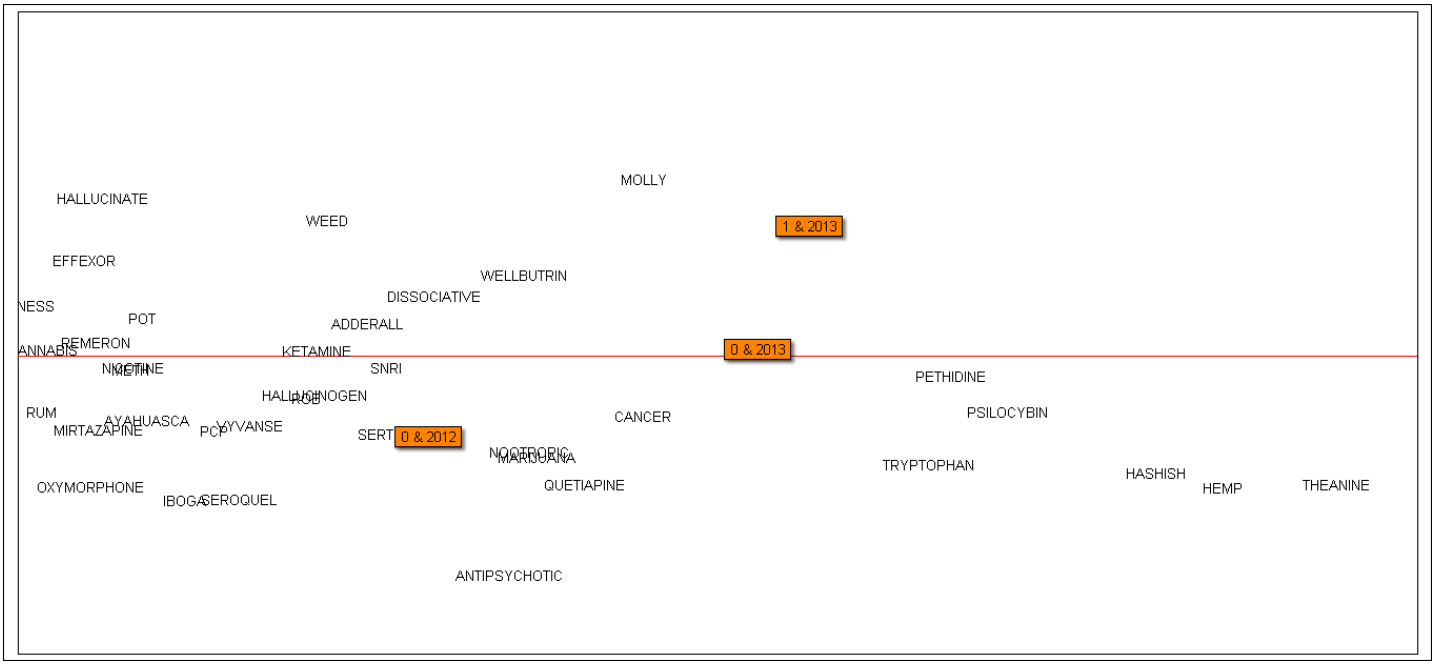


Figure 17b– drug-related keywords most distinctive of posts from new and existing users in particular years, Forum A

Words found in first posts from new users in 2013 appear to be quite distinctive for Forum B, and the terms “concerta” and “fentanyl” in particular are found to be substantially separated from other words more associated with posts from existing users in 2013. Furthermore, it appears that dialogue among existing users in 2013 has not deviated much from language used by both existing and new members in 2012. On the other hand, the language used by newcomers of Forum A does not seem to vary much from terms found in posts from existing users, outside of a slight affinity for the term “molly” associated with the latter.

These findings can also be contextualized by examining posting volume in the manner demonstrated earlier, restricted now to only cases that represent a user’s first post.

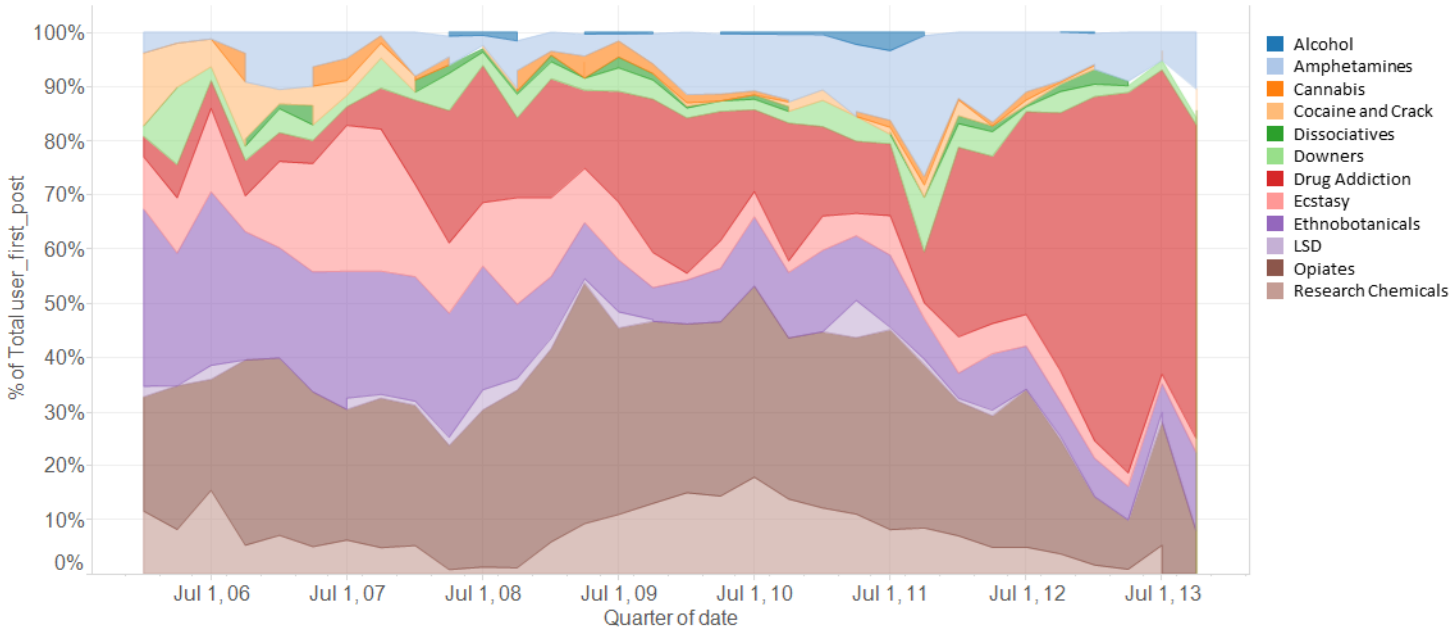


Figure 18a– percent of total first posts over time by section, Forum A

Across new users on Forum A, first posts appear to be concentrated in the Drug Addiction section of the website, mirroring the overall trend in the website's total composition. However, in examining this graph in relation to trends in Forum A's overall volume of posts, it appears that the recent decline in the share of posts related to Research Chemicals may be fueled in part by a diminishing interest in new users; first-time posters in the Research Chemicals section of Forum A appear to have all but vanished.

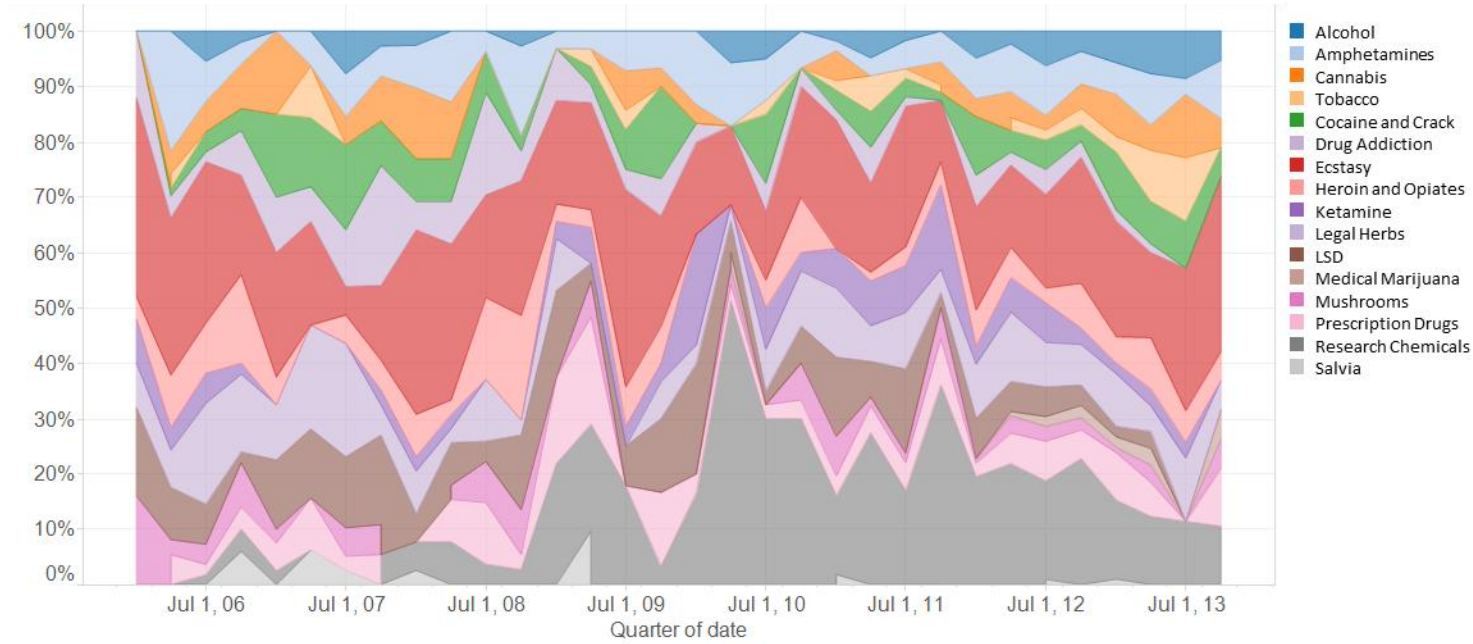


Figure 18b– percent of total first posts over time by section, Forum B

Newcomers to Forum B also appear to be growing less interested in Research Chemicals, demonstrating slow but consistent shift towards posts related to Ecstasy. Among smaller sections of the website, small spikes can be noticed in new users posting about Amphetamines and Prescription Drugs in 2013, along with decreases in the share of new posts dedicated to Tobacco and Alcohol.

CONCLUSION

The analytic methods demonstrated in this report can be expanded and combined to produce far more granular insights than shown here, but this brief exploration of these approaches has already revealed many compelling patterns embedded in the data collected from these two case study forums. The unique social and behavioral information found in these data offers researchers a chance to gain deeper insight into how drug users have been using the internet to discuss illegal substances and how their conversations have been changing in recent years. While an in-depth investigation of these trends is out-of-scope for this proof-of-concept paper, more natural language, social network, and machine learning applications could reveal much about previous and ongoing trends related to substance abuse and public health, and continuing research of this kind could very well serve as a powerful exploratory tool for rapidly identifying emerging trends related to new substances and chemicals, new drug use patterns, new slang terms, early-stage dissemination of information about illegal narcotics, and the spread of misinformation that could become a larger health concern if not quickly addressed.