Measuring the Unmeasurable: Estimating True Population of Hidden Online Communities

Jonah Gibbon
University of Cambridge
jmg229@cantab.ac.uk

Tina Marjanov
University of Cambridge
tm794@cam.ac.uk

Alice Hutchings
University of Cambridge
ah793@cam.ac.uk

John Aston
University of Cambridge
jada2@cam.ac.uk

***Abstract*—The true size of hidden populations is an important aspect when staging interventions or devising policies, yet is inherently difficult to obtain due to its nature. In this paper we present a novel approach for hidden population estimation by leveraging activity measured on underground forums. The proposed method consists of two main components. First, we determine the overlap of populations across forums by evaluating users' behavioural patterns. Subsequently, we employ a Bayesian model tailored for extrapolating data from multiple systems in order to estimate the actual population size. We estimate the true number of people participating in online discussion to be 2-8.5 times higher than observed on major cybercriminal forums, and 1.5-3.5 times higher than observed on extremist forums. Our research contributes to a deeper understanding of fringe populations and offers insights into the potential magnitude of participation in online forums beyond what is readily apparent.**

***Index Terms*—multiple systems estimation, underground forums, cybercrime, extremism**

I. Introduction

Internet forums allow users to openly discuss various topics in an anonymous and relatively consequence-free way. The anonymity of posters encourages users to share their unfiltered opinions, which has led to the rise of many problematic communities participating in (cyber)criminal activities or sharing extremist views. Policy makers, law enforcement, and regulators -as well as academics- often need to monitor these communities to estimate their effects on society, consider interventions, inform policies and coordinate law enforcement effort. While measuring the number of posts or user activity may give us some idea about the number of people involved, such proxies might be problematic and deceptive. Specifically, many of these forums are small and hidden, with users frequently changing sites, making it impossible to monitor all forums. This leads to underestimates for total activity of the population when a substrata of forums is researched. Not knowing the prevalence of problematic views or activities in society has important implications, such as not efficiently allocating resources to prevent harm. In this paper we are interested in estimating the total activity of two types of problematic online communities, namely cybercriminals and extremists.

Unlike most previous research that does not account for undetected activity, we present a method that extrapolates the true population sizes of these hidden communities based on recorded activity across a substrata of prominent online forums. Our method operates in two steps. We first identify users that are active across several forums by observing their behavioural patterns. Specifically, we take into account their username, typical time of posting, and the post content to predict whether a user matches any other account on a different monitored forum. This allows us to calculate the number of unique users that appear across several forums. In the second step, we then extrapolate the number of active users on unmonitored forums, forming estimates for the population size. This is done by applying an existing Bayesian model [1] designed for extrapolating population size of multiple systems data.

It is important to clarify that we do not attempt to estimate the number of users who are not actively posting on these forums (sometimes referred to as *lurkers*), but rather the total number of users actively posting on other hidden forums. The group we estimate will be referred to as the ”hidden” or ”unobserved” population, and the (total) ”population” refers to the total number of users that post on both monitored and hidden forums.

We find that the true number of English speaking users participating on cybercrime-related forums varies significantly over time and is estimated to be between 2-8.5 times larger than evident by simply observing these forums. Similarly, we estimate the true population of English speaking extremists participating on public forums to be 1.5-3.5 times larger than that observed. Notably, we observe a significant increase in the *estimated* extremist population in June 2015 and May 2019, despite no significant increase in the *monitored* population, coinciding with Donald Trump announcing his candidacy for the US 2016 election, as well as spikes in the *estimated* cybercriminal population in June 2018 and later in April 2020. Our findings highlight the importance of actual population estimation, which gives more insightful results compared to simpler participation measurements.

Our main contributions are summarised as follows:

- We present a novel method for hidden population estimation using online forum data.
- We apply our method to two datasets consisting of scraped cybercriminal and extremist underground forums to estimate the undetected population. We estimate the true number of people participating on cybercriminal online forums to be 2-8.5 times higher than that observed,

and 1.5-3.5 times higher than that observed on extremist forums.
- We estimate an increase in the true cybercriminal population shortly after the first COVID-19 pandemic is declared.
- We estimate an increase in the true extremist population around the time Donald Trump announced his presidential candidacy for the US 2016 and 2020 election.
- We find US English spelling is over-represented in both extremist and cybercrime forums.

This paper is structured as follows. In §II the relevant background material is covered, including a brief overview of the development of multiple systems estimation, and a summary of Bayesian statistics. §III introduces the two datasets analysed in this paper; ExtremeBB and CrimeBB. It describes the method used to produce our final estimates, which is broken into two main subsections: the first outlining the method used to form an estimate for the number of active accounts that span different forums, and the second describing the Bayesian model used to estimate the total population size. The results are discussed in §IV and a short conclusion is given in §V. There is an Appendix describing the Bayesian model in greater mathematical detail.

II. Background

A. Underground Cybercriminal and Extremist Forums

A large proportion of online deviant behaviour in some way revolves around various forums and platforms. They provide an anonymous space to discuss beliefs, learn, share knowledge, and trade various services or goods. Several underground forums have become an important hub dedicated to the discussion and promotion of hacking, cheating and scamming techniques, exchanging of illegal services, goods, and stolen data [2, 3]. Additionally, some welcome various extremist views, becoming echo chambers and leading to (further) radicalisation and attacks on perceived enemies of the community [4, 5, 6].

As such, forums provide a wealth of information to researchers. While a significant body of research provides insights into behaviours of various underground communities [2, 7, 8], less is known about their true sizes. Measurements such as the number of participants and the volume of content produced can act as a misleading proxy, usually being an underestimate. Closest to our contribution are works by Cabrero-Holgueras et al. [9] and Vu et al. [10] that devise a methodology to identify related accounts on the basis of content posted and username handles, but do not use the information to extrapolate the number of active users on unmonitored forums.

B. History of Multiple Systems Estimation

Estimating the sizes of hard-to-reach populations -referred to as multiple systems estimation- dates back to 1896, when Petersen used a mark-recapture technique to estimate the sizes of plaice populations [11]. For this mark-recapture method to be accurate, there must be a sufficiently large overlap in the samples taken of the population. While this may be valid in an ecological setting, it is too restrictive when applied to populations that wish to remain hidden, such as those involved in criminal activities.

As a result, there has been recent development in multiple systems estimation when being applied to these 'hidden' populations [12, 13, 14]. The modern adaptations of the capture-recapture and multiple system estimation approach have proven effective for estimating the sizes of hard to observe populations, such as human trafficking victims [15], sexual aggressors [16] or Māori people [17].

However, the computational cost of many of these statistical models scales exponentially with the number of sources of data being used, making them infeasible when analysing data collect across multiple (≥ 6) samples/forums. Manrique-Vallier [1] proposed a Bayesian model that scales linearly with the number of sources of data while producing consistent results to previous models. We build our method around this approach. The model is outlined in §III-C, while a detailed description is provided in the Appendix.

C. Bayesian Theory Overview

Bayesian theory is the field of statistics used for parameter estimation using Bayes' theorem. Let θ denote a parameter to be estimated, and $P(\theta)$ denotes an initial assumption on θ's distribution. After observing data x from a fixed distribution $f(x \,|\, \theta)$, the posterior distribution $P(\theta \,|\, x)$ is proportional to $f(x \,|\, \theta)P(\theta)$. This distribution represents the new likelihood of θ having now observed x.

In hierarchical Bayesian models, that is models that have parameters with a hierarchical dependency, the distribution $P(\theta \,|\, x)$ is unlikely to have a closed form, and thus is difficult to sample from. We can sample from an approximate distribution using a Gibbs Sampling Algorithm (GSA); a type of Markov Chain Monte Carlo method. This involves sequentially sampling from the conditional posterior of each parameter in the model, which usually does have a closed form, to form a Markov chain whose stationary distribution is equal to $P(\theta \,|\, x)$. By running this Markov chain for a sufficiently long time, we can sample from the desired distribution.

III. Methodology

Here we describe our method to estimate the number of English-speaking cybercriminals and extremists active on all public forums. We clarify that this method does not estimate the number of *inactive* users that consume content rather than actively post, and thus when we refer to the "hidden" or "unobserved" population we mean those active on unmonitored forums. We limit the scope of our analysis to two distinct types of online deviant behaviour: cybercrime and extremism, however this method also applies to any other internet forum data. Additionally, we only use data from the most prominent forums; the described method will produce estimates for the total population size including unmonitored, less prevalent public forums.

We first calculate which accounts are active on multiple forums, by comparing 3 features; their usernames, the type of content they post, and the time of day the content is posted, over a selected time window. The values of these windows are chosen to allow for a sufficient overlap of activity across forums, while maintaining accurate results for the time period analysed. Should two accounts be similar in each feature within the time window, they are deemed the same user posting on those two forums. Note that because of the anonymity of posters there is no certainty in these matches, however we justify these conditions and believe they provide good estimates. It is necessary to compute the users whose activity spans multiple forums to apply Manrique-Vallier's [1] model.

Once the number of overlapping active users across different forums is determined, it is then used in the Bayesian model developed by Manrique-Vallier. By assuming the population is a fixed, yet unknown quantity, and that each user is active on some subset of forums independent to any other user (i.e. a Multinomial model), a posterior distribution for the number of active accounts on unmonitored forums can be produced using a GSA. The expectations of these distributions are then combined with the observed accounts to form the final estimates for the total number of active posters on cybercriminal and extremist public forums.

The rest of this section is divided into four parts. Section *A* introduces the two datasets used for the estimation in greater detail. Section *B* details estimating the overlap of users across different forums, including how the username, content and time of activity are used to determine this. Section *C* describes Manrique-Vallier's model in greater detail. Finally, we discuss ethical considerations in Section *D*.

A. CrimeBB and ExtremeBB

We apply our estimation method to two separate datasets – CrimeBB [3] and ExtremeBB [10] – that contain posts scraped from forums dedicated to the discussion of cybercrime-related topics or extremist beliefs, respectively. The two datasets are maintained and shared with researchers by the Cambridge Cybercrime Centre[1]. CrimeBB contains 112 million posts made by 6 million users across 38 cybercrime-related forums. Topics discussed across the forums include hacking, programming, legal and illegal money making methods, malware, trade of various datasets, online game cheating and similar. ExtremeBB contains 57 million posts made by 420 thousand users across 16 extremist forums. The forums discuss inceldom, pickup artistry, looksmaxxing, various conspiracy beliefs, white supremacy, trolling, etc.[2] Each forum generally consists of a number of boards on which registered users can start a thread. In general, only registered users can post content on these threads and participate in discussion.

While posts on certain forums of the datasets date back to 2002, we analyse the time period for which the largest portion of forums are active, specifically between January 2012 and December 2022. Only English speaking forums are considered, since one of the behavioural traits we analyse involves comparing the content posted by accounts. We also limit our analysis to surface websites (i.e. normal websites as opposed to websites on the *dark web*, accessible through the Tor system), to better understand how publicly discussed cybercrime and extremism has evolved in recent years. We leave analysis of dark web for future research. Table I shows the number of posts, accounts and forums for both datasets. The subsets used for our estimation are in bold, while the full dataset is in brackets.

Forum	ExtremeBB	CrimeBB
No. posts	**36M** (57M)	**21M** (112M)
No. accounts	**56k** (420k)	**2.4M** (4.8M)
No. forums	**13** (16)	**18** (38)

TABLE I
CORE DATASET STATISTICS FOR EXTREMEBB AND CRIMEBB

B. Classifying users across forums

Estimating the number of users that span different forums is not straightforward, as users are not required to use identifying characteristic such as a consistent username. Vu et al. [10] devised a method for finding a lower bound for the number of users spanning ExtremeBB, by stipulating that two accounts were the same user if their usernames were the same and sufficiently 'rare', and the time of day they posted were similar. However these conditions are strict and result in an underestimate for the number of users spanning multiple forums. We improve on this method by relaxing these conditions as described below in greater detail. Three metrics were used to uniquely identify the same user that appears across different forums: username handles of accounts, typical timestamp of posts, and typical content of posts.

1) Content Metric: The post content was first cleaned, by removing all hyperlinks, non-alphabetic characters, and stop words in the NLTK stopword corpus [18]. A pre-trained language detection model 'fastText' [19, 20] developed by Facebook AI Research was used to filter any non-English posts from the data. We found that 5.48% of CrimeBB posts and 4.96% of posts for ExtremeBB were non-English. A list of 1,768 words spelt uniquely in UK and US dialects was also identified, and the ratio of the frequency of these words in posts were calculated for each user.[3] If the ratio of US to UK spelt words was less than $1/4$ the user was classified as 'UK', and if it was greater than 4 they were classified as 'US'. Those that did not fall into the above categories were classified as mixed, and those that did not use any of these words were left unclassified. 'UK' and 'US' users were not compared to each other due to the likelihood they were from different regions.

[1] See https://www.cambridgecybercrime.uk.

[2] Inceldom is the paranoid sub-community formed around the inability to find a romantic or sexual partner. Pickup artistry offers advice and training to pick up, date, and have sex with women. Looksmaxxing refers to the techniques used to enhance men's physical attractiveness, including whitemaxxing, the process of someone changing the colour of their skin to appear more white. Trolling refers to online stalking and harassment.

[3] This is not an exhaustive list of all words spelt differently in US and UK dialects.

We believe this ratio confidently classifies those that use UK and US dialects, and any accounts that don't fall into these categories should be investigated in more detail.

A vectorisation natural language processing model [21] was used to embed features from the cleaned content into a 200-length vector. Two separate models were trained for each dataset. This was done in order for the model to recognise the niche terminology used on these platforms [22]. The weighted averages of these embeddings were calculated using the term frequency-inverse document frequency, to encapsulate the average topics discussed by each user. It was deemed that two content vectors were similar if the cosine similarity between them was greater than $1/2$. This threshold was determined through examples, some of which are included in Table II.

Sentence 1	Sentence 2	CrimeBB Similarity	ExtremeBB Similarity
Can you hack this?	I'd like to hack your computer	0.6217	0.824
I voted for UKIP in the last election because I don't like foreigners	Immigrants do not help our country	0.654	0.647

TABLE II
Examples of content posted and their cosine similarity using vectorisation models trained on ExtremeBB and CrimeBB

2) Time Metric: For each account, the timestamps of every post were mapped to the circumference of a unit circle to represent the time of day they were created, with (1,0) representing midnight and (-1,0) representing midday. A von Mises distribution $\pi(\mathbf{x};\boldsymbol{\mu},\kappa)$ was fitted to this data with probability density function proportional to $\exp\left(\kappa\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}\right)$, where $\kappa \geq 0, \mathbf{x}, \boldsymbol{\mu} \in S^1$. Standard maximum likelihood theory gives the estimates for the two parameters as

$$\hat{\boldsymbol{\mu}}^{\mathrm{MLE}} = \frac{\bar{\mathbf{x}}}{R} \quad \text{where} \quad \bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i \quad \text{and} \quad R = ||\bar{\mathbf{x}}||_2$$

and

$$\frac{I_1(\hat{\kappa}^{\mathrm{MLE}})}{I_0(\hat{\kappa}^{\mathrm{MLE}})} = R \tag{1}$$

where $\mathbf{x}_i$ denote the data points on the circle, I_j denotes the jth modified Bessel function and m is the sample size.

Although the estimate for $\boldsymbol{\mu}$ is accurate, the estimate for κ is over confident for small sample sizes, such as $m = 1$ corresponding to a point mass at $\mathbf{x}_1$. Therefore the value of R was rescaled to the new value $\tilde{R} = mR/(m+1)$ as to give an approximate uniform distribution for small sample sizes, small changes in the metric as the sample size increases, and consistent large sample (> 60) results. For computational purposes the approximation for $\hat{\kappa}^{\mathrm{MLE}}$ developed by Sra [23] was used to solve equation (1).

Two fitted distributions X, Y were considered similar if the metric $\frac{D(X\,||\,Y)}{2} + \frac{D(Y\,||\,X)}{2} < 1$, where $D(X\,||\,Y)$ denotes the Kullback-Leibler divergence. This threshold equates to two small sample distributions ($m < 5$) separated by no more than three hours being classified as similar, while imposing the strict condition that users that post regularly ($m > 60$) must be separated by no more than one hour on average to be classified as the same.

3) Name Metric: Liu et al. [24] found that although users across platforms are not obliged to use the same/similar usernames, many choose to do so. As a result, we assume a user active on multiple forums will be operating under similar usernames. We classify two usernames u_1, u_2 as similar if the Jaro similarity string metric [25] of each username in lowercase is greater than 0.9. The Jaro metric lies between 0 and 1, with 1 representing an exact match and 0 representing two strings with no matching characters. This string metric is motivated by its flexibility in transposing characters in the username, as opposed to other types of edit distance metrics that allow substitutions and deletions. The threshold of 0.9 was chosen as to balance finding fuzzy matches but not being overconfident. Examples of the Jaro metric applied to synthetic usernames are presented in Table III.

Username 1	Username 2	Jaro string metric
JohnSmith	johnsmith1	0.967
Batman	Fat man	0.849
Butterfly on your nose	Butterfly on your face	0.909

TABLE III
Jaro similarity string metric for synthetic usernames

4) Scoring and Output: Two users were classified as the same person if all of the above conditions were met. In the cases where two accounts on one forum matched to the same account on a second forum, the score

$$\frac{1-C}{3(1-0.5)} + \frac{1-N}{3(1-0.85)} + \frac{T}{3}$$

was calculated and the username with the highest value was deemed the same (C, T, N denote the three respective metrics). Users active across more than two forums were determined by combining pairwise matches.

C. Multiple Systems Approach

We now describe the model developed by Marique-Vallier, which produces the final estimate for the total population using the number of users spanning K forums. The Appendix provides a more mathematically detailed description.

We first assume N, the size of the total population, is a fixed yet unknown quantity, and each user is active on some randomly chosen subset of forums independent of other users. Therefore we assume that our data comes from a multinomial model, and it suffices to determine the probabilities $f(\mathbf{x}\,|\,\theta)$ of a user being exclusively active in the subset of forums $\mathbf{x}$ (θ denotes a shape parameter). We cannot consider each forum independent, as it is not unreasonable to assume users will be active on forums that share common themes. To address this problem, Manrique-Vallier proposed to fit a Non Parametric Latent Class Model (NPLCM) to the probabilities $f(\mathbf{x}\,|\,\theta)$, a model first described by Dunson and Xing [26].

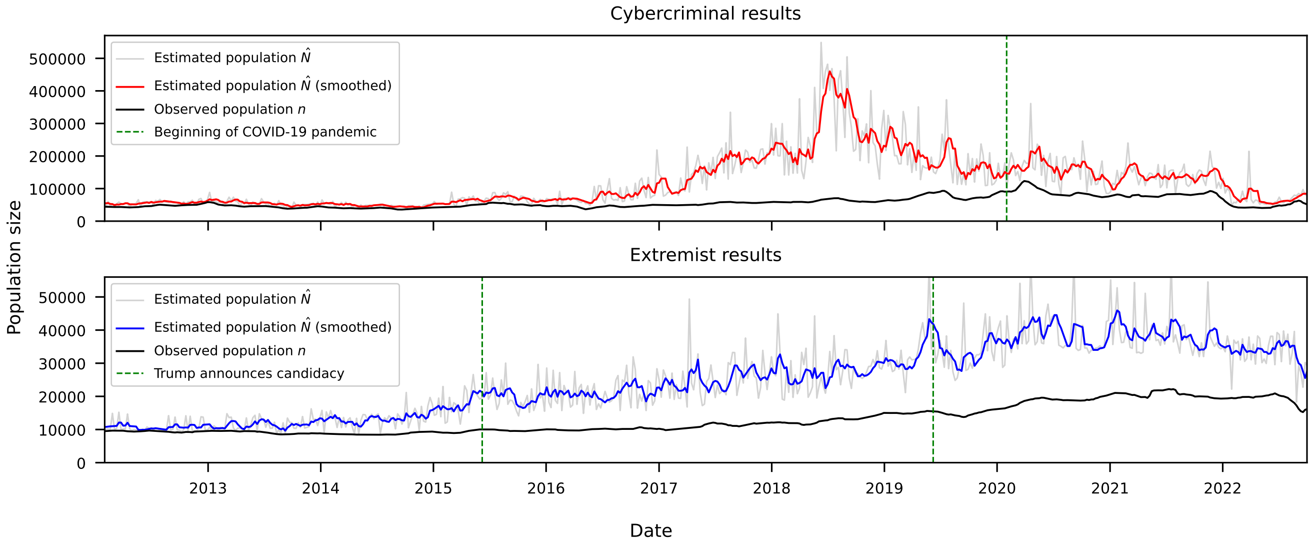

Fig. 1. Time sequential estimates for N using CrimeBB (top) and ExtremeBB (bottom) between January 2012 and December 2022, including a moving average, and the observed accounts n. Note that the vertical axes are not the same.

The NPLCM is defined as a countably infinite mixture of independent product-Bernoulli distributions, and places no restrictive assumptions on $f(\mathbf{x} \mid \theta)$, as proved by Dunson and Xing. This forms the hierarchical Bayesian model

$$\begin{aligned} x_i \mid z &\sim \text{Bernoulli}(\theta_{iz}) \\ z &\sim \text{Discrete}(\mathbb{N}, (\pi_1, \pi_2, \ldots)) \\ \boldsymbol{\theta} := \theta_{ij} &\overset{\text{iid}}{\sim} \text{Beta}(1, 1) \\ \boldsymbol{\pi} := (\pi_1, \pi_2, \ldots) &\sim \text{SB}(\alpha) \\ \alpha &\sim \text{Gamma}(a, b) \end{aligned}$$

where a, b are tuning parameters, iid means independent and identically distributed, x_i represents if the user is active or not in forum i, $\text{SB}(\alpha)$ denotes the stick breaking process with parameter α [27], and $1 \leq i \leq K$, $j \in \mathbb{N}$.

Manrique-Vallier [1] addresses two issues with this approach. First, the countably infinite mixture of product-Bernoulli distributions is replaced with a M finite mixture, where M is chosen to be sufficiently large as to not affect the final results. This exploits the fact that the mixture is sparse, and most of the product-Bernoulli distributions have negligible effect. The second problem is that one of the shape parameters in the NPLCM depends implicitly on the value of N, and hence a GSA would not work. This is rectified by introducing a new set of parameters of which the size does not depend on N, and hence this method is viable. The (improper) prior distribution for N is chosen to be $1/N$, in order for the conditional probabilities of each parameter to be standard distributions.

The first 90% of iterations of the GSA were discarded as a burn-in period, and the last 10% were used to produce independent samples from the posterior, of which the empirical mean $\hat{N}$ and standard deviation $\hat{\sigma}$ were calculated. This average $\hat{N}$ was taken as the final estimate for N.

The GSA took 10000 iterations to reach the stationary posterior distributions for CrimeBB and ExtremeBB, with M equal to 500 and 275 respectively. Choosing a larger value for M saw no change in the final results and hence is sufficiently large. The values for a and b were set to 1 and 1/2, respectively, in order to keep the model sparse.

D. Ethics

This research project was granted ethics approval from the department's ethics committee. The datasets used have been collected from publicly available websites through the use of web scrapers, and informed consent cannot be gained from all members of the forum. However, under the British Society of Criminology's Ethics Statement [28], informed consent may not be required for research into online communities where the data is publicly available, and the research outputs focus on collective rather than individual behaviour. All steps of the analysis are performed in an automated way using various natural language processing, machine learning and statistical tools to minimise the researchers' exposure to the forum content and also preserve the privacy of the posters. Our analysis cannot identify any individuals and we do not attempt to de-anonymise users at any stage of the pipeline.

IV. Results

Our method was run 570 times in order to produce weekly estimates for the total population size for cybercriminals and extremists between January 2012 and December 2022. The time window for comparing online activity was set to 4 weeks (cybercriminals) and 12 weeks (extremists), as to give time sensitive results while still detecting a sufficient overlap across forums. Figure 1 shows the estimated population size of cybercriminals (top) and extremists (bottom) based on the observed accounts measured by the CrimeBB and ExtremeBB datasets respectively. The observed accounts – the number of active accounts within the time window analysed – is marked in black, and the estimated total population is marked in grey. These estimates were smoothed with a 5 week moving average to better understand how the population size has changed, and are marked in red and blue respectively.

A. Cybercriminal Population

The estimates for the population engaging with cybercrime ($\hat{N}$) are up to 8.3 times larger than that of the observed population (n), with an estimated population of up to 550 thousand. We highlight the disparity between the observed and estimated population; the Pearson Correlation Coefficient (PCC) between $\hat{N}$ and n is equal to 0.53. The growth rate of the standard deviation of the estimates exhibited a sub-linear order, approximately $\mathcal{O}\left((N-n)^{0.64}\right)$, suggesting a greater confidence in the estimates when more activity is observed.

Figure 2 shows that up until January 2016, the number of unobserved users per observed user is relatively low. We then see a significant growth in the unobserved population, and thus the estimated population, reaching a global maximum in late 2018 and early 2019. After which the population generally appears to decrease despite little difference in the observed population, however it reaches a local maximum during the COVID-19 pandemic (mid 2020).

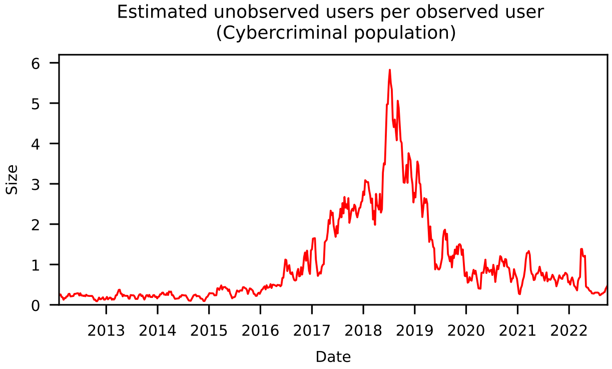

Fig. 2. Ratio of estimated unobserved active cybercriminals to observed active cybercriminals between January 2012 and December 2022.

The large change in the population between 2017 and 2021 may be attributed to users moving across forums regularly, with the majority of users being detected within this time period. This would suggest the cybercriminal population may not have decreased or increased as drastically as shown in Figure 1, but instead we observe a snapshot for the total number of active cybercriminals, before users move to other sites. The small increase in the population following the beginning of the COVID-19 pandemic is likely attributable to users having more free time to participate on these forums.

B. Extremist population

The estimates for the number of extremists ($\hat{N}$) are up to 2.7 times larger than the number of observed users (n), with an estimated total population of up to 54 thousand. The PCC between $\hat{N}$ and n is equal to 0.83, suggesting a larger correlation between the observed and total population of extremists compared to cybercriminals. The standard deviation was found to follow the same trend as with the cybercriminals, growing sub-linearly like $\mathcal{O}\left((N-n)^{0.55}\right)$, again suggesting a higher confidence in estimates when there is more activity.

Figure 3 shows the relationship between the observed and unobserved population is relatively stable over time. We estimate there are up to two unobserved users per each observed user. The number of unobserved users per observed user grows sharply from 2015 to 2016, plateaus until 2020, before gently declining. We want to bring attention to the quick growth in the estimates between 2015 and 2016, coinciding with the US elections at the beginning to 2016. Around the same time as the presidential candidacies were announced, a large jump in extremist participation occurred, and thus we believe this activity was political. The population also increases in mid 2019, coinciding with the run up to the 2020 US elections. Radical movements were linked to these elections, specifically surrounding Donald Trump and the storming of the United States Capitol in 2021, and our estimates provide useful insight to changes in extremist activity before these elections.

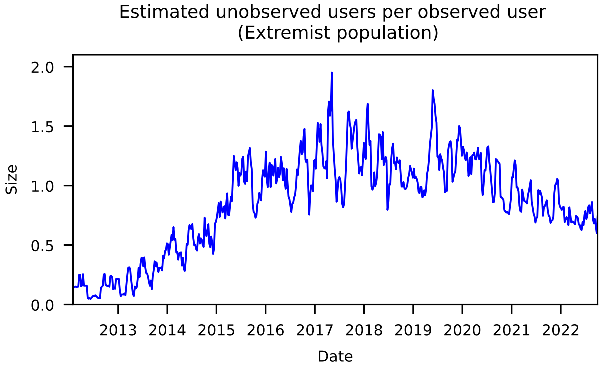

Fig. 3. Ratio of estimated unobserved active extremists to observed active extremists between January 2012 and December 2022.

The gentle decrease in extremist behaviour following 2021 may be due to the relaxation of restrictions following the COVID-19 pandemic. The return to normal may have reduced the frustration of the population and the attraction of extremist forums. Alternatively, extremist views may have became normalised among the population and the conversation moved away from dedicated forums onto more mainstream platforms. More research is necessary to understand these movements.

C. Geographic properties of the population

The CrimeBB and ExtremeBB datasets collect no geographical measurements of where posts are made, and as a result it is difficult to give a description of the locations of those active on these sites. However the measurements taken of the different English spellings give some indication of this (see Table IV). At least one of the distinctly British or distinctly American spelt words needed to appear in the post for it to be classified, hence many accounts were left unclassified. The portion of unclassified users appearing in ExtremeBB is around 40%, while over 90% of CrimeBB posters were left unclassified. We attribute the considerable difference to the nature of posts in the two datasets. Specifically, users posting in extremist forums tend to write more elaborate essay-like posts, while cybercriminals post shorter, sometimes single-word posts and sometimes include code. If we look at only the posts that were successfully classified, we can see that the vast majority in both cases contain American grammar and only a small minority contain British.

	Unknown	UK	US	Mixed	Total
	48k	3k	55k	14k	120k
ExtremeBB	39.6%	2.9%	45.7%	11.8%	100%
	(removed)	4.7%	75.7%	19.6%	100%
	2.09M	7k	203k	11k	2.31M
CrimeBB	90.4%	0.3%	8.8%	0.5%	100%
	(removed)	3.1%	92.0%	4.9%	100%
World Pop. *Inner circle, 2020*		21.7%	70.8%	7.5%	100%

TABLE IV
PROPORTIONS OF US, UK, MIXED AND UNCLASSIFIED DIALECT USING ACCOUNTS ON EXTREMEBB AND CRIMEBB

The proportions do not match the distribution of American and British English across the world, with American English being over-represented in both cases.[4] We speculate this is partly because American English has become the de facto version of English, especially in countries where English is most people's second language, due to the influence of American pop-culture, media and technology. However, even with that in mind, the data indicates North Americans represent the majority of users posting on English-speaking extremist and cybercrime forums.

D. Evaluation

We are unable to empirically check our method for correctness due to the lack of ground truth related to the nature of the problem. Thus we critically examine our method to highlight patterns and perform sanity checking as best we can.

We first investigate the relationship between the estimated population and the main characteristics of observed data during a specific time period. In particular, we are interested in how the number of forums included in the analysis affects the results. The total number of forums actively monitored by the Cambridge Cybercrime Centre has fluctuated in the last decade, and is plotted below in Figure 4. This can be attributed to underground forums being shut down, or users moving to different sites.

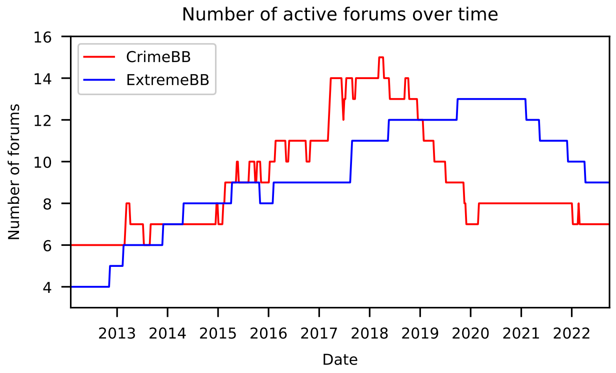

Fig. 4. Number of active forums monitored by Cambridge Cybercrime Centre between January 2012 and December 2022.

To understand this relationship, estimates were calculated after sequentially adding forums into the analysis by decreasing size. These estimates were calculated over the time periods where the most forums were active (March 2018 for cybercriminal population, March 2020 for extremist population), and are presented in Figure 5.

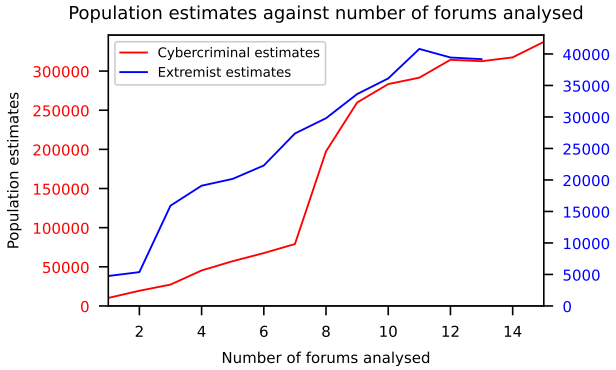

Fig. 5. Effect of number of forums analysed on population estimates for time period March 2020 (extremists) and March 2018 (cybercriminals).

It was found that the cybercriminal population estimates plateau after 9 forums are analysed, whereas the extremist estimates don't plateau until 11 forums were analysed, suggesting our results are more accurate between the time periods February 2015 to December 2019 (cybercriminals) and August 2017 to December 2021 (extremists) where this was the case. The plateaus in Figure 5 suggests that extensively collecting more data from less prevalent forums than those monitored by the Cambridge Cybercrime Centre will not adversely change our estimates.

[4] The distribution between American, British and Mixed English was computed based on the Three Circle Model of World Englishes [29]. The model splits the countries in three circles: the inner (native speakers), the outer (non-native, but English is a part of a country's chief institutions) and expanding circle (English is taught as a foreign language and not part of a country's chief institutions). We consider the countries of the inner circle, which include USA, UK, Canada, Australia, New Zealand and Ireland. While several modern models or extensions have been proposed to complement or expand upon it, the model remains widely recognised in the field of sociolinguistics.

Consistent with previous work [10] are our findings on topic overlaps between the various extremist beliefs. Specifically, we find several clusters of forums with shared common topics (e.g. trolling, nationalism, inceldom) within ExtremeBB, as well as overlap between these clusters. We are unable to find such defined clusters on CrimeBB. This is likely due to the less specialised nature of many of the forums, where people may discuss general and cybercrime-related topics as well as exchange tools, services and data. More overlap between certain common topics – and consequently the populations – among the extremist forums leads to fewer unobserved members, further leading to smaller estimated population. The opposite is true among the cybercriminal population. This finding is particularly interesting since the number of forums analysed at any point in time is relatively comparable between the two datasets (see Figure 4), yet the ratios between the estimated and observed populations are drastically different.

We believe another potential explanation for the difference is a broader landscape that is available for discussion of certain extremist beliefs, compared to cybercriminal topics. For example, extremist conversations might also happen on alternative platforms, such as Gab and Parler, or on other more general social media sites.

Finally, w use the number of registered users reported by forums as a *proxy* for true number of users and compare that to the estimated values. We note that users reported on the forums include inactive accounts and ”lurkers”, which do not perfectly match the profile of estimated users, which we earlier defined as *active* participants on the forums. Regardless, they represent a portion of unobserved population and are as such the closest observable proxy for the full unobserved population. As we do not have the data for the reported population over time, we instead look at the number of users reported by the forum at a single point in time (February 2023).

Table V shows the ratios between active population as measured in CrimeBB/ExtremeBB and population reported by forums themselves, where available. The ratios among the cybercrime-related forums range between 2.7 and 7.4, while the ratios of extremist forums range between 1.7 and 2.9. In general, the pattern is consistent with the ratios between the observed and estimated population. Specifically, the ratios are higher and more spread out in CrimeBB, and lower and less spread out in ExtremeBB.

V. Conclusion

We presented a novel method to estimate hidden population size and applied it to two distinct domains; cybercriminal and extremist populations. We find that estimated populations fluctuate significantly over time and can reach several times that of the observed population. We produce estimations that are not visible by eye and in some cases counter-intuitive, giving important conclusions that typical surface-level statistics overlooks. However there are several possible extensions and improvements that could be made.

We emphasise again that our scope is narrow, as our method is only capable of estimating population that fits within the defined constraints. Specifically, we are only able to make estimations about the population that is *active* on *public* forums and communicates in *English*. We excluded non-English activity due to a lack of ready-available tools and models for processing. Analysing additional non-English accounts may provide greater insight into the geographical locations of users beyond the Anglosphere, giving us a more complete picture of the populations. We focus on the publicly visible interactions due to their availability, however it is believed many interactions will occur privately. Thus our estimates are likely a lower bound for the population that engages in cybercriminal activity or shares extremist beliefs.

	Site	Reported	Active	Ratio
CrimeBB	kernelmode.info	13k	2k	6.5
	mpgh.net	4000k	572k	7
	hackforums.net	5495k	738k	7.4
	blackhatworld.com	1382k	435k	3.2
	cracked.io	4340k	1044k	4.2
	breachforums.is	104k	36k	2.9
	nulled.to	5158k	1897k	2.7
ExtremeBB	stormfront.org	380k	179k	2.1
	looksmax.org	33k	15k	2.2
	pick-up-artist-forum.com	190k	65k	2.9
	kiwifarms.hk	105k	84k	1.25
	incels.is	25k	15k	1.7

TABLE V
Comparison between number of members reported by the forum and number of members measured in CrimeBB and ExtremeBB

Next, we perform some sanity checking and explore the relationship between the number of forums in the dataset and estimations. However, the lack of ground truth makes it very hard to evaluate the results so more work needs to be done in order to fully understand how the estimations depend on other measurement-related factors. Cross validating the results using other multiple systems estimation methods [13] would be one way of strengthening our conclusions.

Lastly, present work serves as a proof of concept for the method and only captures a very coarse picture, merging all the forums within a time frame. Future work could extend the method to track movements of users across forums over time. This is particularly interesting for cybercriminal forums that are often taken down by law enforcement and users migrate over to a new or existing platform. Naturally, the method is not limited to only forums. Alternatively, the method could be expanded to social networks and other platforms, both on surface and dark web, where memberships are less defined to better understand the number of people actively engaging with an idea (e.g. hashtag, keywords) and estimate the population engaging in an activity or sharing a certain belief.

Finally, knowing the true size of the population is crucial for policymakers, law enforcement, technology companies and the broader community. First, it allows more effective monitoring of the right communities. Next, it allows for more efficient resource allocation in order to prioritise interventions based on the scale of the issue and ensure that adequate funding and personnel are allocated. Additionally, once interventions have been put in place, monitoring the real changes in sizes of the community helps evaluate their effectiveness. Overall, we believe our results highlight the importance of true population estimation, on which policies and interventions should be based.

Acknowledgements

We thank the Cambridge Cybercrime Centre for access to the CrimeBB and ExtremeBB datasets, as well as our colleagues for their valuable feedback. This work is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme, grant No 949127 (for JG, TM & AH). TM is supported by King’s College, Cambridge and the Cambridge Trust. As part of the open-report model followed by the Workshop on Attackers & CyberCrime Operations (WACCO), all the reviews for this paper are publicly available at https://github.com/wacco-workshop/WACCO/ tree/main/WACCO-2024.

References

[1] Daniel Manrique-Vallier. “Bayesian population size estimation using Dirichlet process mixtures”. In: *Biometrics* 72.4 (2016), pp. 1246–1254.

[2] Jack Hughes and Alice Hutchings. “Digital drift and the evolution of a large cybercrime forum”. In: *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. 2023, pp. 183–193.

[3] Sergio Pastrana et al. “Crimebb: Enabling cybercrime research on underground forums at scale”. In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 1845–1854.

[4] Catharina O’Donnell and Eran Shor. ““This is a political movement, friend”: Why “incels” support violence”. In: *The British Journal of Sociology* 73.2 (2022), pp. 336–351.

[5] Anh V Vu, Alice Hutchings, and Ross Anderson. “No Easy Way Out: The Effectiveness of Deplatforming an Extremist Forum to Suppress Hate and Harassment”. In: *arXiv preprint arXiv:2304.07037* (2023).

[6] Bruce Hoffman, Jacob Ware, and Ezra Shapiro. “Assessing the threat of incel violence”. In: *Studies in Conflict & Terrorism* 43.7 (2020), pp. 565–587.

[7] Mayur Gaikwad et al. “Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools”. In: *Ieee Access* 9 (2021), pp. 48364–48404.

[8] Marti Motoyama et al. “An analysis of underground forums”. In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. 2011, pp. 71–80.

[9] José Cabrero-Holgueras and Sergio Pastrana. “A methodology for large-scale identification of related accounts in underground forums”. In: *Computers & Security* 111 (2021), p. 102489.

[10] Anh V Vu et al. “Extremebb: Enabling large-scale research into extremism, the manosphere and their correlation by online forum data”. In: *arXiv preprint arXiv:2111.04479* (2021).

[11] Carl Georg Johannes Petersen. “The yearly immigration of young plaice in the Limfjord from the German sea”. In: *Rept. Danish Biol. Sta.* 6 (1896), pp. 1–48.

[12] Sheila M Bird and Ruth King. “Multiple systems estimation (or capture-recapture estimation) to inform public policy”. In: *Annual Review of Statistics and Its Application* 5 (2018), pp. 95–118.

[13] Bernard W Silverman. “Multiple-systems analysis for the quantification of modern slavery: classical and Bayesian approaches”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 183.3 (2020), pp. 691–736.

[14] Lax Chan, Bernard W Silverman, and Kyle Vincent. “Multiple systems estimation for sparse capture data: Inferential challenges when there are nonoverlapping lists”. In: *Journal of the American Statistical Association* 116.535 (2021), pp. 1297–1306.

[15] Maarten Cruyff, Jan Van Dijk, and Peter GM van der Heijden. “The challenge of counting victims of human trafficking: Not on the record: A multiple systems estimation of the numbers of human trafficking victims in the Netherlands in 2010–2015 by year, age, gender, and type of exploitation”. In: *Chance* 30.3 (2017), pp. 41–49.

[16] Martin Bouchard and Patrick Lussier. “Estimating the size of the sexual aggressor population”. In: *Sex offenders: A criminal career approach* (2015), pp. 349–371.

[17] Peter GM Van Der Heijden et al. “Multiple system estimation using covariates having missing values and measurement error: Estimating the size of the Māori population in New Zealand”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.1 (2022), pp. 156–177.

[18] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O’Reilly Media, Inc.”, 2009.

[19] Armand Joulin et al. “Bag of Tricks for Efficient Text Classification”. In: *CoRR* abs/1607.01759 (2016). arXiv: 1607.01759. URL: http://arxiv.org/abs/1607.01759.

[20] Armand Joulin et al. “FastText.zip: Compressing text classification models”. In: *CoRR* abs/1612.03651 (2016). arXiv: 1612.03651. URL: http://arxiv.org/abs/1612.03651.

[21] Piotr Bojanowski et al. “Enriching Word Vectors with Subword Information”. In: *arXiv preprint arXiv:1607.04606* (2016).

[22] Jack Hughes and Alice Hutchings. “Argot as a Trust Signal: Slang, Jargon & Reputation on a Large Cybercrime Forum”. In: *Workshop on the Economics of Information Security (WEIS)*. 2023.

[23] Suvrit Sra. “A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of I s (x)”. In: *Computational Statistics* 27 (2012), pp. 177–190.

[24] Jing Liu et al. “What's in a name? An unsupervised approach to link users across communities”. In: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, pp. 495–504.

[25] Matthew A Jaro. “Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida”. In: *Journal of the American Statistical Association* 84.406 (1989), pp. 414–420.

[26] David B Dunson and Chuanhua Xing. “Nonparametric Bayes modeling of multivariate categorical data”. In: *Journal of the American Statistical Association* 104.487 (2009), pp. 1042–1051.

[27] J.E. Griffin and M.F.J. Steel. “Stick-breaking autoregressive processes”. In: *Journal of Econometrics* 162.2 (2011), pp. 383–396. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2011.03.001.

[28] British Society of Criminology. *Statement of Ethics*. 2015.

[29] Braj B Kachru. “World Englishes and English-using communities”. In: *Annual review of applied linguistics* 17 (1997), pp. 66–87.

[30] Robert J Connor and James E Mosimann. “Concepts of independence for proportions with a generalization of the Dirichlet distribution”. In: *Journal of the American Statistical Association* 64.325 (1969), pp. 194–206.

Appendix

Here we include the mathematical details for the model used in §III-C, including the derivation of the GSA.

A. Notation

Let N denote the total population size, and $L = \{1, 2, \ldots, K\}$ be the set of forums. Let $\mathcal{P}(L)$ be the power set of L and n_S be the number of users active on forums $S \in \mathcal{P}(L)$ and not active on forums $\mathcal{P}(L)\backslash S$. It is sufficient to estimate $n_\emptyset$, the number of unobserved cases, since the total population size $N = n_\emptyset + \sum_{\emptyset \neq S \in \mathcal{P}(L)} n_S$. Let $n = N - n_\emptyset$ denote the total number of active users being measured.

Let $(\mathbf{x}_i)_{i=1}^N$ denote which intersection of forums each user lies in. This is done by denoting $\mathbf{x}_i \in \{0,1\}^K$, where $(\mathbf{x}_i)_j = 1$ if user i is active on forum j, and 0 otherwise. Notice there is the natural bijection between $\{0,1\}^K$ and $\mathcal{P}(L)$, and hence the abuse of notation $\mathbf{x}_i = S$ will be used to denote $(\mathbf{x}_i)_j = 1 \iff j \in S$.

B. The Multinomial Model

Suppose N is a fixed yet unknown quantity, and each user is randomly assigned to some intersection of forums $\mathbf{x} \in \mathcal{P}(L)$ with probability $f(\mathbf{x}\,|\,\theta)$, where θ is some shape parameter. Then the probability of observing the data $\mathcal{N} := (n_S)_{\emptyset \neq S \in \mathcal{P}(L)}$ would be the multinomial distribution

$$\binom{N}{n} f(\mathbf{0}\,|\,\theta)^{N-n} \prod_{\substack{\emptyset \neq S \in \mathcal{P}(L) \\ \mathbf{x}=S}} f(\mathbf{x}\,|\,\theta)^{n_S} \; \mathbf{1}(N \geq n)$$

where $\mathbf{1}$ denotes the indicator function. It suffices to specify $f(\mathbf{x}\,|\,\theta)$ to fully determine the model. This is where the Non Parametric Latent Class Model (NPLCM), developed by Dunson and Xing [26], is fitted. It is defined as the hierarchical model

$$\begin{aligned} x_i \,|\, z &\sim \text{Bernoulli}(\theta_{iz}) \\ z &\sim \text{Discrete}(\mathbb{N}, (\pi_1, \pi_2, \ldots)) \\ \boldsymbol{\theta} := \theta_{i,j} &\overset{\text{iid}}{\sim} \text{Beta}(1,1) \\ \boldsymbol{\pi} := (\pi_1, \pi_2, \ldots) &\sim \text{SB}(\alpha) \\ \alpha &\sim \text{Gamma}(a,b) \end{aligned}$$

where $\text{SB}(\alpha)$ denotes the stick breaking process, $1 \leq i \leq K$, $j \in \mathbb{N}$ and a, b are predetermined constants.

The stick breaking process is a Dirichlet process used to draw a random infinite-discrete distribution of decreasing probabilities, such that there is sparsity away from the first few probabilities. The process is simple: for each $i \geq 1$, take $\beta_i \overset{\text{iid}}{\sim} \text{Beta}(1,\alpha)$ and set $\pi_i = \beta_i \prod_{j=1}^{i-1}(1-\beta_j)$. This is equivalent to taking a $\text{Beta}(1,\alpha)$ random position on a unit length stick, and marking this position as π_1. Then discarding the stick to the left of π_1, mark an iid position on what is to the right of π_1 and call this $\pi_1 + \pi_2$. By repeating this, an infinite random sequence of positive numbers is formed such that they sum to 1, as well as having the majority of their probabilities concentrated in the first few terms. Notice the smaller the value of α, the greater β_i will be on average, and thus the quicker π_i will converge to zero (almost surely) and the more concentrated the resultant distribution will be in the first few terms.

With this new model we can fully specify the distribution $P(\mathcal{N}\,|\,\boldsymbol{\theta}, \boldsymbol{\pi}, N)$. It is equivalent to marginalising $P(\mathcal{N}, \mathbf{z}\,|\,\boldsymbol{\theta}, \boldsymbol{\pi}, N)$ over $\mathbf{z} = \{z_i^S \in \mathbb{N} : S \in \mathcal{P}(L), 1 \leq i \leq n_S\}$, where $P(\mathcal{N}, \mathbf{z}\,|\,\boldsymbol{\theta}, \boldsymbol{\pi}, N)$ equals

$$\begin{aligned} &\frac{N!}{(N-n)!} \prod_{i=1}^{N-n} \pi_{z_i^\emptyset} \prod_{j=1}^{K} (1 - \theta_{jz_i^\emptyset}) \times \\ &\prod_{\substack{\emptyset \neq S \in \mathcal{P}(L) \\ \mathbf{x}=S}} \left\{ \frac{1}{n_S!} \prod_{i=1}^{n_S} \pi_{z_i^S} \prod_{j=1}^{K} (\theta_{jz_i^S})^{x_j} (1 - \theta_{jz_i^S})^{(1-x_j)} \right\} \end{aligned} \tag{2}$$

multiplied by the indicator function $\mathbf{1}(N \geq n)$. This final model is a robust conclusion, as it imposes no assumptions on the distribution of $f(\mathbf{x}\,|\,\theta)$, as well as being in the appropriate format for Bayesian estimation, as it comprises only of multiplying simple functions together. The following section discusses the appropriate alterations to this model to effectively compute the posterior distribution.

C. Alterations to the NPLCM

The first problem when sampling from the model in equation (2) is that $\boldsymbol{\pi}$ is infinitely dimensional. However, $\boldsymbol{\pi}$ is sparse and so we may assume that the probabilities become negligible for sufficiently large i. Therefore $\mathbf{z}$ can be viewed as taking values on the truncated set of natural numbers from 1 to M. The exact value of M is determined through trial and error; increasing the value until the posterior distribution is stable. The values of $\boldsymbol{\pi}$ will be determined by truncating the stick breaking process at $i = M - 1$ and normalising π_M.

The second problem is that the size of $\mathbf{z}$ depends implicitly on the value of N, and so cannot be conditioned on when implementing a GSA. Note that $\mathbf{z}^+ := \{z_i^S : \emptyset \neq S \in \mathcal{P}(L), 1 \leq i \leq n_S\}$ can be conditioned on as it places no assumptions on N, and the remaining variables $\mathbf{z}^\emptyset := \{z_i^\emptyset : 1 \leq i \leq n_\emptyset\}$ can be replaced with a new set of variables, the number of which do not depend on N. This is done by defining $\boldsymbol{\omega} \in \mathbb{N}^M$ with ω_k denoting the number of unobserved cases where $z_i = k$. The importance of this is it derives a new representation of the model $P(\mathcal{N}, \mathbf{z}^+, \boldsymbol{\omega} \,|\, \boldsymbol{\pi}, \boldsymbol{\theta}, N)$ where the number of variables is fixed, namely

$$\begin{aligned}&\binom{N}{n, \omega_1, \ldots, \omega_M} \prod_{m=1}^{M} \left(\pi_m \prod_{k=1}^{K} (1 - \theta_{km})\right)^{\omega_m} \times \\ &\prod_{\substack{\emptyset \neq S \in \mathcal{P}(L) \\ \mathbf{x}=S}} \left[\frac{1}{n_S!} \prod_{i=1}^{n_S} \pi_{z_i^S} \prod_{j=1}^{K} (\theta_{j z_i^S})^{x_j} (1 - \theta_{j z_i^S})^{(1-x_j)}\right] \times \qquad (3) \\ &\mathbf{1}\left(\sum_{m=1}^{M} \omega_m = N - n\right)\end{aligned}$$

This is the model from which the GSA is derived from.

D. The Gibbs Sampling Algorithm

Many of the variables in model (3) are independent of one another, and hence their conditional distributions are standard. However care is needed for the cases $\boldsymbol{\pi}$, N and $\boldsymbol{\omega}$.

To derive the conditional posterior $P(\boldsymbol{\pi} \,|\, \ldots)$ the stick breaking process must be revisited.[5] Previously $\pi_i = \beta_i \prod_{j=1}^{i-1}(1 - \beta_j)$, where each $\beta_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ for $1 \leq i \leq M - 1$. Suppose instead that the β_i were drawn from independent Beta(a_i, b_i) distributions where the parameters were allowed to change. Connor and Mosimann [30] derived that, should this be true, then $P(\boldsymbol{\pi} \,|\, \ldots)$ follows a generalised Dirichlet distribution $\mathcal{GD}(\mathbf{x} \,|\, \mathbf{a}, \mathbf{b})$ defined as

$$\begin{aligned}&\left[\prod_{i=1}^{M-1} B(a_i, b_i)\right]^{-1} x_M^{b_{M-1}-1} \times \\ &\prod_{i=1}^{k-1} \left[x_i^{a_i - 1} \left(\sum_{j=i}^{M} x_j\right)^{b_{i-1} - (a_i + b_i)}\right]\end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^{M-1}$ such that $\mathbf{x} \geq 0$ and $||\mathbf{x}||_1 \leq 1$, $x_M := 1 - \sum_{i=1}^{M-1} x_i$, $B(\cdot, \cdot)$ denotes the Beta function and $(\mathbf{a})_i = a_i, (\mathbf{b})_i = b_i$.

It is easy to check that if $a_i = 1 + c_i$ and $b_i = \beta + \sum_{j=i+1}^{M} c_j$, where c_i are constants, then

$$\mathcal{GD}(\mathbf{x} \,|\, \mathbf{a}, \mathbf{b}) = \mathcal{GD}(\mathbf{x} \,|\, \tilde{\mathbf{a}}, \tilde{\mathbf{b}}) \prod_{i=1}^{M-1} x_i^{c_i}$$

where $\tilde{a}_i = 1$ and $\tilde{b}_i = \beta$. Therefore if we choose $c_i = n_i + \omega_i$ where n_i is the number of z_i^S that equal k, then this gives the correct posterior for $\boldsymbol{\pi}$.

Since $\sum_{i=1}^{M} \omega_i = n_\emptyset$, sampling from the posterior $P(N \,|\, \ldots)$ is not possible as $\boldsymbol{\omega}$ fully specifies N. Sampling from the joint conditional distribution $P(N, \boldsymbol{\omega} \,|\, \ldots)$ avoids this problem, which is proportional to

$$P(N) \frac{N!}{\omega_1! \ldots \omega_M!} \rho_1^{\omega_1} \ldots \rho_M^{\omega_M} \times \mathbf{1}(N = n + n_\emptyset)$$

where $P(N)$ is the prior chosen for N, $\rho_i := \pi_i \prod_{j=1}^{K}(1 - \theta_{ji})$ and $n_\emptyset = \sum_{i=1}^{M} \omega_i$. Should $P(N) \propto 1/N$ then the above is proportional to

$$\begin{aligned}&\binom{n + n_\emptyset - 1}{n_\emptyset} \left(\sum_{i=1}^{M} \rho_i\right)^{n_\emptyset} \left(1 - \sum_{i=1}^{M} \rho_i\right)^{n} \times \\ &\frac{n_\emptyset!}{\omega_1! \ldots \omega_M!} \rho_1^{\omega_1} \ldots \rho_M^{\omega_M} \left(\sum_{i=1}^{M} \rho_i\right)^{-n_\emptyset}\end{aligned}$$

This is the product of a negative binomial distribution NB$(n, 1 - \sum_{i=1}^{M} \rho_i)$ and a multinomial distribution Multi$(n_\emptyset, (p_1, \ldots, p_M))$ where $p_i \propto \rho_i$. Therefore drawing from $P(N, \boldsymbol{\omega} \,|\, \ldots)$ can be done by sequentially drawing from $n_\emptyset$ and then $\boldsymbol{\omega}$.

With the conditional distributions established, Algorithm 1 describes the GSA for the altered NPLCM. The initial state of the algorithm was set to the expectation of each parameter under their prior distribution, and $n_\emptyset$ initially being set to n since its prior was improper.

Algorithm 1 One iteration of the Gibbs sampling algorithm for altered NPLCM described in equation (3).

```
Sample $z_i^S \sim$ Discrete$(\{1, \ldots, M\}, (p_1, \ldots, p_M))$, where $p_i \propto \pi_i \prod_{j=1}^{K} (\theta_{ji})^{x_j} (1 - \theta_{ji})^{1-x_j}$ and $\mathbf{x} = S$.
Sample $\theta_{jk} \sim$ Beta$(n_{jk} + 1, n_k - n_{jk} + \omega_k + 1)$, where $n_k$ is the number of $z_i^S = k$, and $n_{jk}$ is the number of $z_i^S = k$ where $j \in S$.
Sample $\beta_k \sim$ Beta$(1 + c_k, \alpha + \sum_{i=k+1}^{M} c_i)$ for $k < M$ and $\beta_M = 1$, where $c_k = n_k + \omega_k$.
Set $\pi_k = \beta_k \prod_{i<k}(1 - \beta_i)$.
Sample $\alpha \sim$ Gamma$(a - 1 + M, b - \log \pi_M)$.
Sample $n_\emptyset \sim$ NB$(n, 1 - \sum_{i=1}^{M} \rho_i)$ where $\rho_i = \pi_i \prod_{j=1}^{K}(1 - \theta_{ji})$.
Set $N = n + n_\emptyset$.
Sample $\boldsymbol{\omega} \sim$ Multinomial$(n_\emptyset, (p_1, \ldots, p_M))$, where $p_i \propto \rho_i$.
```

This algorithm is a computationally efficient method to estimate N. Each iteration of the algorithm involves sampling $\mathcal{O}(KM)$ times from standard distributions, making it highly scalable in the number of forums K. This method was tested on a large set of data in the paper [1] by Manrique-Vallier, and thus we have not tested the theoretical limitations of the model.

[5] The $\ldots$ denote all other variables