Anonymity and Identity Online*

Florian Ederer† Paul Goldsmith-Pinkham‡ Kyle Jensen§

October 25, 2023

Abstract

Economics Job Market Rumors (EJMR) is an online forum and clearing house for information about the academic job market for economists. It also includes content that is abusive, defamatory, racist, misogynistic, or otherwise "toxic." Almost all of this content is created anonymously by contributors who receive a four-character username when posting on EJMR. *Using only publicly available data* we show that the statistical properties of the scheme by which these usernames were generated allows the IP addresses from which most posts were made to be determined with high probability.[1] We recover 47,630 distinct IP addresses of EJMR posters and attribute them to 66.1% of the roughly 7 million posts made over the past 12 years. We geolocate posts and describe *aggregated* cross-sectional variation—particularly regarding toxic, misogynistic, and hate speech—across sub-forums, geographies, institutions, and IP addresses. Our analysis suggests that content on EJMR comes from all echelons of the economics profession, including, but not limited to, its elite institutions.

JEL Codes: C55, D83, D91, L86, Z13

Keywords: cryptography, internet privacy, large language models, toxic speech

*We are grateful to Jason Abaluck, Tim Bresnahan, Nicholas Christakis, Damon Clark, Forrest Crawford, Rachael Meager, Barry Nalebuff, Matt Notowidigdo, Michael Ostrovsky, Julia Simon-Kerr, Steve Tadelis, Catherine Tucker, Mahboud Zabetian, seminar audiences at the NBER Summer Institute and Yale, and several economists who wish to remain anonymous for helpful comments. The analysis described in this manuscript was allowed to proceed by the Yale HRPP, IRB protocol ID 2000034072. Our study uses only publicly available pages on EJMR, the same pages crawled and indexed by major search engines and does not access any non-public pages or hidden URLs. Due to the presence of misinformation about our paper, we compiled an FAQ in Appendix A.

†Boston University Questrom School of Business, ECGI, and NBER, florian.ederer@gmail.com

‡Yale School of Management and NBER, paul.goldsmith-pinkham@yale.edu

§Yale School of Management, kyle.jensen@yale.edu

[1]The scheme changed on May 17, 2023 after the first release of the abstract of this paper on May 16, 2023. The scheme changed again on July 21, 2023 after the first public presentation of this paper on July 20, 2023.

"Economics is what economists do."

– Jacob Viner, quoted in Spiegel (1987, p. 814)

1 Introduction

Economics Job Market Rumors (http://www.econjobrumors), henceforth EJMR, is an anonymous internet message board featuring discussion about economics, the economics profession, and, in particular, the annual job market for PhD economists. However, the discussion board is active year-round and EJMR users also post much content unrelated to economics. The site is popular: SimilarWeb estimates that in early 2023 EJMR received 2.5 million visits per month with an average of 6.45 pages viewed per visit. In comparison, the same figures for the websites of the National Bureau of Economic Research (NBER) and the American Economic Association (AEA) are 1.1 million and 991,000 visits and 2.09 and 2.76 pages per visit, respectively.

EJMR is controversial in the economics profession. It has been called "a breeding ground for personal attacks of an abusive kind" (Blanchard, 2017), a "cesspool of misogyny" (Romer, 2017), "4chan" and "4chan for economists" (Lowrey, 2022; Taylor, 2023), and "evidence of a toxic environment for women in economics" (Wolfers, 2017).[2] A substantial fraction of EJMR consists of content that is abusive, defamatory, racist, misogynistic, or otherwise "toxic." Such content exists in spite of both automated and manual moderation. Examples include "the whole point of women is to get railed and make babies" and "The biggest enemies of America are: Blks" and "America lost its war against blks [...] At least until we resolve to final solution" and "University of Stupid Chinese" and "The average woman has a 15% smaller brain than the average man" and "the fastest route to a qje is to grift and be black." Comments also target particular individuals, including examples such as "Should Jennifer Doleac be executed for her anti-Chinese hatred?" and "Anya Samek [...] took advantage of her initial postdoc position organizing conferences [...] and handling requests for grant proposals to steal ideas" and "Are Vrinda and Hampole in a secret same-sekhs love-hayte relationship?".

In this paper, we show that EJMR's contributors post from locations, institutions, and universities that are intrinsically linked to the academic economics community, including the upper echelons of academia, government, and the private sector. This result runs counter to

[2]The Committee on the Status of Women in the Economics Profession (CSWEP) of the AEA (https://www.aeaweb.org/about-aea/committees/cswep/statement) also condemned "the sexist, racist, homophobic and anti-Semitic statements that have appeared on the Economics Job Market Rumors (EJMR) site, and particularly the harassment and abuse targeted at particular scholars." More than 1,000 signers urged the AEA in 2017 to create a moderated, well-functioning site to provide up-to-the-minute job market information (https://www.iaffe.org/petition-aea-ejmr/). However, the forum EconSpark and the information website EconTrack, both created by the AEA, were largely failures and languish without much use.

the common claim that EJMR is not representative of the economics profession and that its the most frequent users on EJMR are not actually economists. We also confirm that EJMR is sexist, as first documented by Wu (2018) and Wu (2020) and provide new evidence that toxic speech is pervasive across contributors and institutions, including elite universities, suggesting that EJMR perpetuates and amplifies existing inequalities in the economics profession (Bayer and Rouse, 2016; Antecol et al., 2018; Lundberg and Stearns, 2019; Dupas et al., 2021; Hengel, 2022).

EJMR used a specific algorithm to assign usernames to posts written by anonymous contributors. We show that the properties of this algorithm do not anonymize the contributors of posts, but instead allows the IP address from which each post was made to be determined with high probability. To recover IP addresses from the observed usernames on EJMR, we employ a multi-step procedure. First, we develop GPU-based software to quickly compute the SHA-1 hashes used for the username allocation algorithm on EJMR. In total, we compute almost 9 quadrillion hashes to fully enumerate all possible IP combinations and to check which of the resulting substrings of hashes match the observed usernames. For each post, this procedure narrows the set of possible IP addresses from 2^{32} to roughly 2^{16}. Second, we measure which IP addresses occur particularly often in a short time window and use the uniformity property of the SHA-1 hash to test whether these IP addresses appear more often than would likely occur by chance.

Our statistical test minimizes the probability of falsely assigning an IP address to a post because the p-value thresholds we employ are of the order of approximately 10^{-11}. Our approach has several further advantages. First, even though there are 7,098,111 posts on EJMR, we assign exactly *zero* posts to the large set of unaddressable (bogon) IP addresses from which no traffic should legitimately appear on the public Internet. Second, our assignment method precisely pinpoints when EJMR changed its hashing algorithm and assigns exactly *zero* posts if an incorrect hashing algorithm is used. Third, our methods correctly identify the location of users' IP addresses. IP addresses post during the standard work and day time hours of their geolocation and the dominant non-English language of the country of origin of the IP address is the country's native language. Despite this conservative approach our procedure recovers 47,630 distinct IP addresses of EJMR posters and assigns 66.1% of the roughly 7 million total posts to these IP addresses.

We then describe *aggregated* features of posting behavior on EJMR. Despite several attempts to limit its influence, EJMR remains popular. In 2022, the platform averaged 70,000 monthly posts. This allows us to recover roughly 1,100 unique IP addresses per month. Although EJMR has been a popular website for economists since its inception, posting and engagement on the site have surged since the start of the COVID-19 pandemic, especially

in the United States where posting rates tripled. This increase was primarily driven by a very large increase in off-topic forum posts that are mostly unrelated to the job market for academic economists. Off-topic posts in other countries such as Canada, the United Kingdom, Hong Kong, Australia, Germany, Italy, and France also increased, but the increases were less pronounced and more temporary.

Based on the geographic location of the IP addresses we identify, we show that the majority of posts come from large cities in the US (Chicago, New York, Philadelphia) and outside the US (Hong Kong, London, Montreal, Toronto). Smaller US cities with leading research institutions such as Cambridge and Berkeley are also towards the top of the list. Beyond the city and country, the IP addresses we recover also allow us to identify the associated internet service providers. Based on this information, we show that posting on EJMR is pervasive throughout the economics profession. Over 10% of posts originate from university networks including, but not limited to, all top-ranked universities in the United States. A substantial number of posts also come from government agencies, companies, and non-profit organizations employing economists as well as universities around the world.

We then study the distribution of problematic content on EJMR and show that it is widespread among both residential and university IP addresses, and not concentrated among just a small set of contributors. To do so, we deobfuscate the text of EJMR posts and use state-of-the-art transformer models to show that roughly 10% of posts contain text labeled as toxic, 2.5% as misogynistic, and 2.5% as hate speech. Posts marked as toxic, misogynistic, and hate speech are more likely to originate from residential IP addresses than from IP addresses associated with universities, but the difference is small. We also show that IP addresses are very often posting in topics that contain toxic content, even when their own content is not toxic. However, the share of problematic content on EJMR is high even by the standards of anonymous online forums. We show that EJMR posts are more likely to be labeled as toxic, misogynistic, and hate speech than 69, 73, and 95 percent of the 1,000 most popular subforums (subreddits) of Reddit, a popular discussion website where users can post relatively anonymously. We also find that EJMR posts mentioning women are substantially more likely to contain hate speech, misogyny, and toxicity. Moreover, mentions of women attract even more misogyny and hate speech on EJMR than on Reddit, but the increase in toxicity is similar.

2 Methods

2.1 Relationship between IP Addresses and Usernames

The vast majority of EJMR users do not log into the EJMR website using a persistent username of their own selection. Instead, the site uses a scheme it described as follows: "EJMR allows you to post anonymously whereby your post enters the database without a record of personally identifiable information like your IP or email address. However to prevent users from voting for themselves and to help users maintain the same 4 letter indentity [sic] within a thread one way encryption is used to create a 4 letter identity. This is a combination of random strings and the user's IP address which is one way encrypted and then sliced up to create a 4 letter ID which is stored in the database."

Shortly after introducing this scheme EJMR's pseudonymous administrator "Kirk" wrote "...for example you can see this post is also from me by looking at the fddf2 on the left. But I'll give you a million US$ if you can guess my ip."[3] That IP address—twelve years ago—was almost certainly 188.220.40.122.[4]

How we can make such a statement? In brief, we correctly guessed the scheme by which EJMR assigned usernames for most of its history. The statistical properties of that scheme allow us to back out IP addresses from usernames for about 65% of posts on EJMR. We likely identify 100% of the IP addresses that were consistently active on the site and, in thousands of cases, IP addresses active on the site for as little as a week's time. *All of this is achieved using only publicly available data.* With the exception of the aforementioned "million-dollar" address, we will not disclose IP addresses associated with EJMR posts in this paper.

Before we describe how to map usernames to IP addresses, we provide a short introduction to a few concepts that may be unfamiliar to readers trained in economics. First, IPv4, or Internet Protocol version 4, is the prevailing method of addressing devices on the internet. An IPv4 address is a 4-byte, or 32-bit, number in the interval $[0, 2^{32})$, or $[0, 4294967296)$. For example, 2189728028 is a valid IP address. However, this IP address would more commonly be written in the so-called "dotted decimal" notation as 130.132.153.28, wherein each number represents one of the four bytes, or four octets, in the binary representation of the number 2189728028. Each octet is an integer in the range $[0, 256)$. Blocks of IP addresses are assigned by the Internet Assigned Numbers Authority to "autonomous systems:" smaller networks of computers administered by internet service providers (ISPs), government agencies, corpora-

[3] As this post shows, EJMR briefly experimented with five-letter usernames.

[4] This IP address is part of a large block of consumer IP addresses which likely changes hands frequently, could be used by millions of devices, and gives little geographic detail beyond being in or proximate to London. This location is, however, consistent with an article on EJMR and "Kirk" in the German newspaper WirtschaftsWoche in 2011 (https://amp2.wiwo.de/politik/economics-job-market-rumors-die-geruechtekueche-der-volkswirte/5971044.html).

tions, and universities. For example, Yale University owns autonomous system number 29 (AS29) and owns two blocks of 2^{16} addresses (about 130,000 in total) including the above address 130.132.153.28.

Autonomous systems allocate their IP addresses to devices on their networks using a variety of mechanisms. For example, an IP address can be statically assigned to a device, in which case the device can have the same IP address for years. Or, IP addresses can be dynamically assigned to devices through methods like Dynamic Host Configuration Protocol (DHCP), as might be the case on a university wireless network. Devices using DHCP retain IP addresses for minutes to months. Also, devices can be behind "network address translation" or NAT, a method that allows multiple devices to share the same "external" IP address while using a unique "local" address. "Carrier grade" NAT is particularly common for mobile networks (Livadariu et al., 2018). Multiple devices can also share the same IP address when using a proxy, a virtual private network (VPN), or an "anycast" domain name system (DNS) configuration. The latter of these is rare for consumer devices, however, proxies and VPNs are quite common.

As this description should make clear, *IP addresses are not people*. Humans using devices, such as those posting to websites, change IP addresses. They use multiple devices. And, quite often, masses of people can share the same IP address. Therefore, nothing in our manuscript should be construed as identifying *persons*.[5]

Having covered IP addresses, a short description of EJMR's organization is necessary to understand how IP addresses fit into its anonymity protocol. EJMR is built on bbPress (Wordpress Foundation, 2023), which is a version of the popular WordPress blogging software that is customized to host forums. bbPress websites, such as EJMR, are organized by *topics*. Each topic has a URL and a title. For example, the topic at URL `https://www.econjobrumors.com/topic/dream-job-imf-economist` has the title "Dream job: IMF economist". Topics also have *posts*. When a user creates a new topic, that user simultaneously creates the first post. Subsequently, other users can add posts to the topic. Each post has some user-written content, a timestamp, and a username identifying the user who made the post.

As we described previously, most EJMR users do not log into the site using a self-selected username as do users on a typical bbPress site. Instead, the overwhelming majority of posts on the site are made by anonymous users for whom EJMR *generates* usernames. Consider the topic "Dream job: IMF economist". The initial poster asks about how best to gain employment at the IMF and was assigned username 270b. Shortly thereafter, EJMR assigned username

[5]The ability to link several posts coming from the same IP address reveals additional information about the persons posting content on EJMR and in some cases can allow for the identification of specific individuals. However, such an exercise is not the focus of this paper.

dd86 to the user replying "Your country must be a real craphole for IMF to be your dream job."

As EJMR's notification says, the site assigns contributors a *persistent* username for each topic using the contributor's IP address. That is, a contributor commenting on topic t from IP address a always receives the same username, regardless of date, comment content, or browser state, including user-agent and cookies. If, at some later date, user dd86 contributed EJMR from the same IP address and offered more advice in the IMF topic, they would retain the username dd86. However, this contributor will, with probability approaching one, receive a *different* username when they post on a different topic or from a different IP addresses.

How are usernames generated on EJMR? The four-letter usernames comprise solely the characters a-f and 0-9, which suggested to us that the usernames are hexadecimal encoded numbers. Hexadecimal—or base-16—encoding is compact way to write large numbers as alternative to base-10. Hexadecimal digits include the base-10 Arabic digits 0-9 and A, B, C, D, E, and F which represent 10, 11, 12, 13, 14, and 15, respectively. In base-10, the username of the IMF poster 270b is the number $2\times16^3+7\times16^2+0\times16^1+11\times16^0 = 9995$. Hexadecimal is common encoding by which to represent the output of "hash" functions. We guessed that hashing is the technique to which EJMR referred when saying "the user's IP address ... is one way encrypted."[6] Hash functions—such as the SHA and MD5 family of functions—are functions that map data of arbitrary or large range to a domain of fixed size. For example, imagine a hash function f that takes a sentence of text and returns an integer representing the index of the first letter of text in the English alphabet. The sentences are of arbitrary size and the domain is of finite size: $[1-26]$. This would be a bad hash function for most practical uses of hash functions. As the EJMR notice referenced above describes, hashes are "one-way" functions: the output is easily determined from the input, but given an output it is difficult or impossible to know the input. For example, knowing that a sentence begins with "E" does not tell you the sentence. Most cryptographic hash functions do not output small numbers like 26, but rather exceedingly large numbers that are more compactly written in hexadecimal than decimal format. For example, the output of the SHA-1 hash is an integer in the range $[0, 2^{160})$.

The topic pages for websites built on BBPress each contain two identifiers that we thought might be inputs to the username scheme. For example, consider the EJMR topic page `https://www.econjobrumors.com/topic/dream-job-imf-economist`. This topic has a "slug" which is the string "dream-job-imf-economist" and it also has a numeric ID, which has the value 227259 in this case. Topic IDs on BBPress websites are auto-incrementing integer primary keys in the underlying MySQL relational database used by BBPress and WordPress. Topics can be

[6] Strictly speaking, hash functions are not a form of encryption, which is by definition two-way.

accessed by these IDs online. For example, visiting https://www.econjobrumors.com/topic/227259 redirects to https://www.econjobrumors.com/topic/dream-job-imf-economist, showing the same content to visitors.

We guessed that EJMR's usernames were generated as follows

$$u = S(\mathcal{H}(M(t, a, o))) \tag{1}$$

where t is a topic identifier, a is a visitor's IP address, and o is some other data, typically a "salt" which is used to improve the security of data obfuscated by hashing (Ferguson et al., 2010, p. 304). M is a function that mixes together the inputs, including possibly a "stretch", which improves security (Ferguson et al., 2010, p. 304). S is the function EJMR uses by which the output would be "sliced up to create a 4 letter ID". $\mathcal{H}$ is a hash function (Ferguson et al., 2010, p. 77). Because the EJMR usernames are in hexadecimal, we suspected S was a simple function of $\mathcal{H}$'s output.

We had in our possession three different EJMR usernames for which we knew both the topic ID and the IP address from which the post was made. This gave us three concordant sets of u, t, and a. We suspected a was restricted to IPv4 addresses, which are more commonly used than IPv6 addresses. Later, we verified that EMJR's webserver does not respond to IPv6 internet traffic, only IPv4 traffic. Therefore, each EJMR user has an IPv4 address. We observed u and t on EJMR. We began a search for o, M, $\mathcal{H}$ and S with simple guesses by which we attempted to recreate our three observed values of u for our three sets of u, t, and a.[7] Our search was short. We presumed that there was no salt and thus set o to null. We guessed that M was either the concatenation of t and a as strings, or a and t. We furhter guessed that S was a function that merely returned a substring of the hash. Finally, we guessed that the hash function $\mathcal{H}$ was a common hash function such as MD5, SHA-1/224/256/384/512, or CRC32.

We found that o is indeed null. S is the string concatenation of t and a, where a is in the dotted decimal notation. $\mathcal{H}$ is the SHA-1 hash and M returns characters 10-13 of the hexadecimal hash (1-based indexing). That is, if a user visits EJMR from the IPv4 address 131.111.5.175 and posts on the topic with id 227259, EJMR assigns the username c2b1. This is the four character interval at position 10-13 in e8b5eae32c2b197a0ac4cb889a9bbb8f417f3bff which is the hexadecimal encoding of the SHA-1 hash of the string "227259131.111.5.175" (ASCII encoded). In other words, the EJMR username is the hexadecimal representation of the two bytes of data beginning at the 40th bit of the 20-byte big-endian SHA-1 hash. In

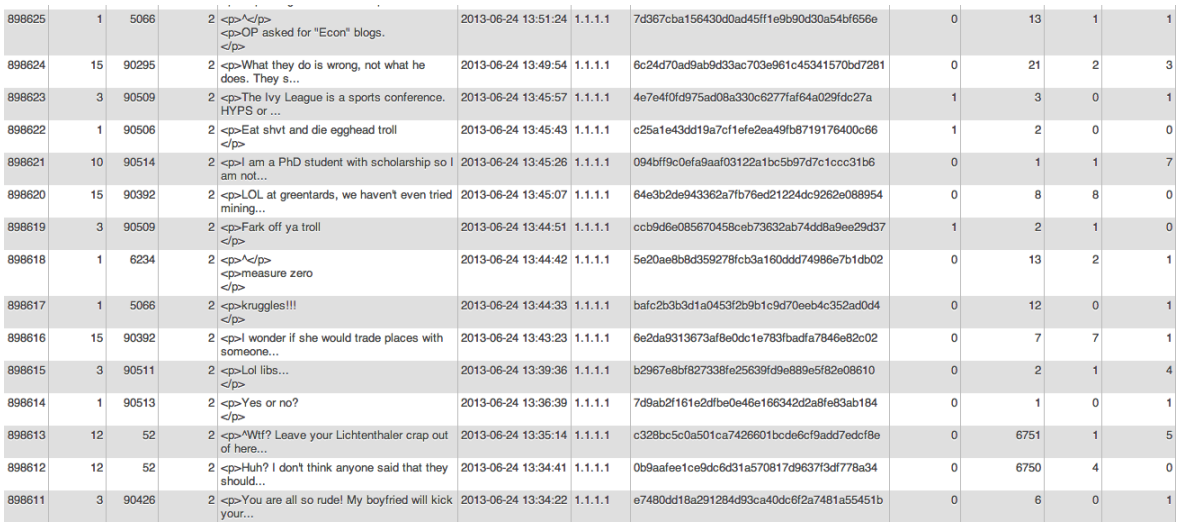
[7] In total, we posted five times on EJMR: three times to verify the hashing scheme and two times to produce a brief set of videos to document how the hashing scheme worked. These videos can be viewed at https://www.youtube.com/watch?v=on5YCsEhGrY&list=PLWWcL1M3lLloToQOE_j1Ys8dQZlckMIIp.

plain English, EJMR combines a visitor's IP address with an integer topic id, hashes that with SHA-1, and uses a part of that hash as the username.

The above is an accurate description of the EJMR username scheme for the period from July 8, 2013 to May 17, 2023. Because this scheme is no longer in use, a skeptical reader may rightly ask how they can verify this claim. We have three responses. First, the results that follow will, in numerous ways, show statistical patterns that would be nigh impossible if we were incorrect about the username generation scheme. Second, on February 7, 2023, we recorded a brief video in which we show how an EJMR username could be computed *prior to posting on the site* with a knowledge of a topic ID and one's IP address. This video can be viewed at https://www.youtube.com/watch?v=on5YCsEhGrY&list=PLWWcL1M3lLloToQOE_j1Ys8dQZlckMIIp. Third, and most importantly, our claim is supported by the public posts of EJMR's administrator. On July 3, 2013, the site administrator wrote "Here is a direct screenshot of all of the fields for each post http://i.imgur.com/1htoXw7.png". This post can be viewed in the WayBack Machine[8] and the screenshot appears in Figure 1. The screenshot shows 15 rows, one for each of 15 different posts. Each post has 12 columns. Column 3 is the topic ID and column 8 shows the SHA-1 hash of a topic ID and the IP address from which the post was made. Readers can easily verify that there is *one and only one* IPv4 address that, when pre-pended with the topic ID, produces the SHA-1 hash shown in the screenshot. For example, there is only a *single* IPv4 address ending in ".42" that, when pre-pended with 6234, produces the SHA-1 hash 5e20ae8b8d359278fcb3a160ddd74986e7b1db02.

At the time this screenshot was shared on EJMR, the website saved the entire SHA-1 hash for each post, but *displayed* just characters 10-13 from the hash. In response to some EJMR user criticism, on July 8, 2013, the site administrator began storing just positions 10-13 in the database instead of the whole hash. The site administrator also elected to purge old hashes from the database. But, for old posts (i.e., posts before July 8, 2013) EJMR began showing positions 9-12 of the SHA-1 hash of each topic-IP pair, as shown in Figure 2. This was mostly likely due to an error on the part of the administrator. BBPress is built with the PHP language, which uses zero-based indexing. So, the PHP code for the EJMR username scheme looks something like `substr(sha1($topic_id . $user_ip), 9, 4)`, which means "take the substring from position 9 for 4 characters" where position 9 is the *tenth* character in the hash because the first character is *zero*. It seems likely that, in an effort to discard whole SHA-1 hashes, the site administrator issued a MySQL command like `update posts set the_sha1_hash = substring(the_sha1_hash from 9 for 4)`. However, MySQL uses 1-based indexing instead of 0-based indexing. Therefore, the effect of this command would be to "shift" the username left for posts made before July 8, 2013. Having issued this SQL

[8] https://web.archive.org/web/20230531180223/https://econjobrumors.com/topic/kirk-31#post-913648



898625	1	5066	2	^ OP asked for "Econ" blogs.	2013-06-24 13:51:24	1.1.1.1	7d367cba156430d0ad45ff1e9b90d30a54bf656e	0	13	1	1
898624	15	90295	2	What they do is wrong, not what he does. They s...	2013-06-24 13:49:54	1.1.1.1	6c24d70ad9ab9d33ac703e961c45341570bd7281	0	21	2	3
898623	3	90509	2	The Ivy League is a sports conference. HYPS or ...	2013-06-24 13:45:57	1.1.1.1	4e7e4f0fd975ad08a330c6277faf64a029fdc27a	1	3	0	1
898622	1	90506	2	Eat shvt and die egghead troll	2013-06-24 13:45:43	1.1.1.1	c25a1e43dd19a7cf1efe2ea49fb8719176400c66	1	2	0	0
898621	10	90514	2	I am a PhD student with scholarship so I am not...	2013-06-24 13:45:26	1.1.1.1	094bff9c0efa9aaf03122a1bc5b97d7c1ccc31b6	0	1	1	7
898620	15	90392	2	LOL at greentards, we haven't even tried mining...	2013-06-24 13:45:07	1.1.1.1	64e3b2de943362a7fb76ed21224dc9262e088954	0	8	8	0
898619	3	90509	2	Fark off ya troll	2013-06-24 13:44:51	1.1.1.1	ccb9d6e085670458ceb73632ab74dd8a9ee29d37	1	2	1	0
898618	1	6234	2	^ measure zero	2013-06-24 13:44:42	1.1.1.1	5e20ae8b8d359278fcb3a160ddd74986e7b1db02	0	13	2	1
898617	1	5066	2	kruggles!!!	2013-06-24 13:44:33	1.1.1.1	bafc2b3b3d1a0453f2b9b1c9d70eeb4c352ad0d4	0	12	0	1
898616	15	90392	2	I wonder if she would trade places with someone...	2013-06-24 13:43:23	1.1.1.1	6e2da9313673af8e0dc1e783fbadfa7846e82c02	0	7	7	1
898615	3	90511	2	Lol libs...	2013-06-24 13:39:36	1.1.1.1	b2967e8bf827338fe25639fd9e889e5f82e08610	0	2	1	4
898614	1	90513	2	Yes or no?	2013-06-24 13:36:39	1.1.1.1	7d9ab2f161e2dfbe0e46e166342d2a8fe83ab184	0	1	0	1
898613	12	52	2	^Wtf? Leave your Lichtenthaler crap out of here...	2013-06-24 13:35:14	1.1.1.1	c328bc5c0a501ca7426601bcde6cf9add7edcf8e	0	6751	1	5
898612	12	52	2	Huh? I don't think anyone said that they should...	2013-06-24 13:34:41	1.1.1.1	0b9aafee1ce9dc6d31a570817d9637f3df778a34	0	6750	4	0
898611	3	90426	2	You are all so rude! My boyfried will kick your...	2013-06-24 13:34:22	1.1.1.1	e7480dd18a291284d93ca40dc6f2a7481a55451b	0	6	0	1

Figure 1: Screenshot of the EJMR MySQL database posted by the administrator "Kirk."

command, without IP addresses or a database backup, it would be impossible to "correct" the usernames. Using the WayBack Machine, readers can verify that the EJMR usernames before May 17, 2023 were as we claim here. For example, the above-mentioned post with hash 5e20ae8b8d359278fcb3a160ddd74986e7b1db02 in Figure 1 had username 8d35 when captured by the WayBack Machine in 2015. This is position 9 to 12 in the hash because this post was among those "shifted" left.

The discussion above should make it clear that the EJMR username scheme was sitting in plain sight for the past decade. Further, the scheme is unsophisticated. It contains no salt (o), stretching function (M), or cipher (non-trivial S). The scheme accomplishes the objective of ensuring that return visitors receive a consistent username in each topic. However, it is an *exceedingly bizarre* choice if one wishes to obfuscate the IP address from which a post originates. The most common use of cryptographic hashes like SHA-1 is to *identify* content. For example, SHA hashes are used to identify content in the git version control system and in the Bitcoin blockchain. The username scheme that EJMR uses nakedly advertises roughly 16-bits of information from each post's 32-bit IPv4 address origin. Because of the trivial choices for M, S, and in particular o (no salt), that information is readily recoverable as we describe in the following section.

2.2 Mapping Usernames to IP Addresses

In order to answer basic questions about the distribution of toxic speech on EJMR, we wished to map EJMR posts to their origin, both in IP address space and geographically. This presented a few challenges. To understand those challenges and how we overcame them,

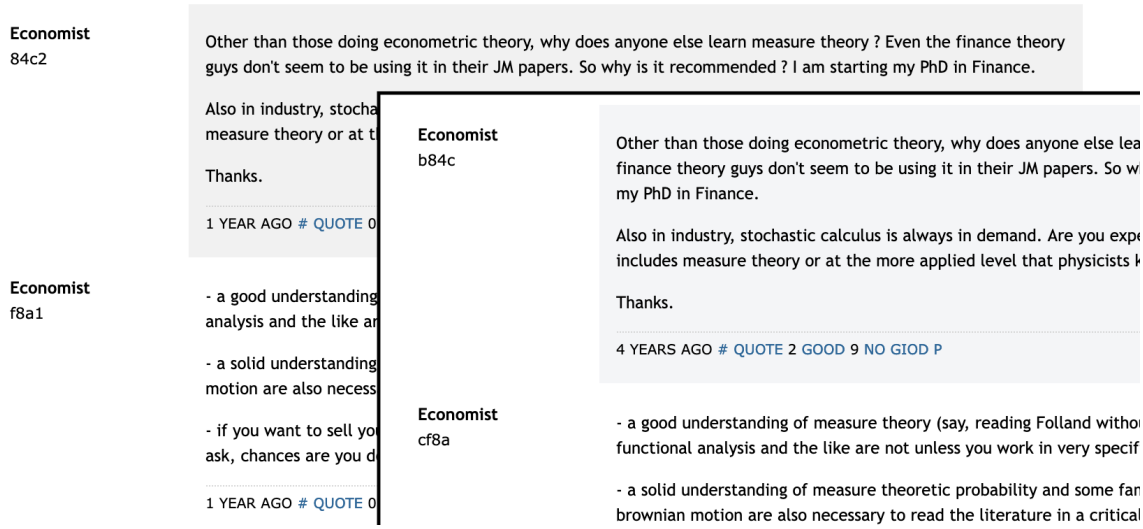


Figure 2: Screenshot of an EJMR topic before and after July 8, 2013, from the WayBack Machine. The leftmost posts shows usernames created using positions 10, 11, 12, and 13 in the SHA-1 hash of the topic id, IP address combination. The inset on the right shows the same posts after July 8, 2013 where the usernames are constructed using positions 9, 10, 11, and 12 of the same hash. We believe this change to be a result of administrator error.

it is helpful to offer some background on the SHA-1 hash. The SHA-1 hash was created by the United States National Institute of Standards and Technology in 1995 (Standard, 1995). Like its predecessor MD5 and its successor SHA-256, the SHA-1 hash uses the so-called Merkle–Damgård construction (Damgård, 1990). Each of these hashes is widely used. For example, as mentioned above, the version control system `git` uses SHA-1 to identify source code changes Spinellis (2012) and SHA-256 is used in the proof-of-work system of Bitcoin (Nakamoto, 2008). These hashes each have a desirable property called the "avalanche" effect whereby small changes in the input produce large changes in the output (Motara and Irwin, 2016). Ideally, a one bit change in the input causes each output bit to flip with probability 0.5. The avalanche effect leads to the "uniformity" property of these hashes such that inputs map uniformly to the output domain. In the case of SHA-1 this means that inputs, such as the topic-IP concatenation from EJMR, are mapped uniformly over the range $[0, 2^{160})$.

The uniformity property of SHA-1 implies that every hash value in the output range is generated with roughly the same probability (Cormen et al., 2022). Of course, the EJMR username scheme uses on a two-byte interval of the SHA-1 hash. To verify that the SHA-1 hash is also uniform over these bytes, we conducted two experiments. First, we chose a topic ID at random and computed the EJMR username for all IPv4 addresses. Second, we chose a random IPv4 address and computed the EJMR username for all extant EJMR topic

IDs. In both cases we found the SHA-1's uniformity to be preserved over the two-bytes used for the EJMR username. Or, more formally, we could not reject the hypothesis that it is uniform using a chi-squared statistic. Therefore, based on this uniformity, given a post with a username, we expect roughly $2^{32}/2^{16} = 2^{16} = 65,536$ IP addresses to have been possible origins of the post (i.e., hashes of the topic-IPv4 concatenations where position 10-13 match with the username).

To see why this is helpful, let us return to our example in Section 2.1 in which the IP address 131.111.5.175 posting on the topic "Dream job: IMF economist" with the topic ID 227259 was assigned the username c2b1. There are exactly 65,028 IPv4 addresses that, when prepended with topic ID 227259, create a hexadecimal SHA-1 hash with c2b1 on position 10-13. Because of the uniformity property of SHA-1 these matching IP addresses are uniformly distributed over the entire IPv4 address range from 0.0.0.0 to 255.255.255.255.

Now imagine the same IP address 131.111.5.175 posts on the topic "Are US-based journals biased towards US data?" with the topic ID 227279. In this case it would be assigned the username 91c2. There are exactly 65,635 IP addresses that would also receive the username 91c2. But there is only *one* IP address, the true IP address 131.111.5.175, which appears in the two sets. As this example illustrates, "true positive" IP addresses—those from which posts on EJMR were actually made—*stick out* because these IP addresses explain more observed usernames on the site than *false positive* IP addresses.

Clearly, in order to figure out the "true" IP address for an EJMR post, we need to know the roughly 65k candidate IP addresses. But, determining those is not trivial because the hash is one-way. One cannot determine the input to a SHA-1 hash given the output (much less a fraction of that output). The only feasible way to determine which approximately 65k IP addresses could have produced a username is to compute the username for each of the 2^{32} IPv4 addresses. This is conceptually easy but intolerably slow in practice. Given that there are 695,364 topics on EJMR and there are 2^{32} IPv4 addresses, we need to perform about 2.98×10^{15} (or, in words, 2.98 quadrillion) SHA-1 hashing operations and then check which of the computed hashes correspond to topic-username combinations observed on EJMR. This computation requires only a handful of lines in high-level languages like Python. However, our initial tests suggested that such an effort would take over 60 years on a single core of a typical modern CPU. Fortunately, this manner of computation is made easier by graphical processing units (GPUs) which are essentially massively parallel computers, but which require specialized programming frameworks such as CUDA.

After obtaining every topic-username combination observed on EJMR (of which there were 5,184,896 in our data set) and developing software that runs on Nvidia GPUs, we determined the IP addresses from which each of the topic-username combination could have originated.

The heart of the software is based on an open-source implementation of the SHA-1 algorithm designed for Nvidia devices from the Mochimo Cryptocurrency project (Mochimo Cryptocurrency Engine, 2023). Some aspects of our task allowed for optimizations that substantially sped up this task. First, because the topic ID is *prepended* to the IP address before hashing, the SHA-1 algorithm could be "primed" once for each topic and fed to each compute core of the GPU. Second, because the IPv4 addresses are merely 32-bit integers, they could be enumerated on the GPU device rather than passed in as strings, thereby limiting GPU-CPU data transfer, which can otherwise be a bottleneck. Third, we required only one pass over the 2^{32} IPv4 addresses for each topic. Roughly speaking, our algorithm passes the GPU a "primed" SHA-1 hash and a list of the observed usernames for a particular topic. Each core of the GPU considers a single IP address and checks if that IP address would produce a username that is observed. If so, the username-IP pair is appended to a list of results that is ultimately passed back to the CPU for output. This process repeats until all 2^{32} IPv4 addresses are checked for a single topic.

The IP enumeration task is "embarrassingly parallel" because topics can be enumerated independently. In the end, this task took about 240 hours (i.e., ten days) of total computing time on Nvidia A100 devices which each have 6,912 cores (Choquette et al., 2021) that operate in parallel. We used multiple devices so the actual time was significantly lower. The device we used—A100 GPUs with 40g of memory—retail for roughly $8,000 at this time. These devices can also be rented hourly. For example, the P4d instances on Amazon's EC2 contain eight A100 devices. Our study could be reproduced for roughly $1,000 on an AWS P4d instance at the hourly on-demand price although, as we describe later, we repeated this hash inversion in triplicate for the purpose of our statistical analysis. Furthermore, because our method used only approximately 0.5Gb of memory per GPU device, it can almost certainly be reproduced with older, less expensive CUDA devices. The intermediate output of the hash inversion required roughly 3Tb of storage space. That said, the statistical analysis of the output of this process also required computers with at least 100Gb of RAM. The source code for this enumeration task is available at `https://github.com/to-be-determined` in the "cuda-sha" directory.

2.3 Probabilistic Identification of IP Addresses

Recall that there are 2^{32} possible IPs a in the space of IPv4 addresses and that there are 16^4 possible EJMR usernames. In Section 2.2 we showed how we generate all possible a in IPv4 space that are associated with each (t, u) observed on EJMR. Because of the uniformity property of SHA-1 each (t, u) observation is equally likely to have been created by any one of

roughly 65,536 different IP addresses.[9]

Although this significantly narrows the set of possible IP addresses, it does not allow us to exactly assign the true IP address associated with a topic-username observation. However, because EJMR prepends the topic ID to the IP address before hashing, the same IP address will be associated with a different hash and thus be assigned a different username when posting on a different topic.

Formally, consider the statistical problem of determining whether an IP address is "randomly" in the set of approximately 65,536 a assigned to each (t, u). We are able to make progress on this by examining how often that IP address shows up in the *other* (t, u) combinations across the site (specifically, in other topics t). In our data, we observe m distinct topics t, and for every topic t, we have m_t distinct usernames. Hence, there are a total of $n = \sum_{t=1}^{m} m_t$ distinct topic-username combinations.[10] Additionally, let $A_{(t,u)}$ denote the set of roughly 65,536 a associated with (t, u), U_t denote the set of observed usernames for a topic t, and let n_a denote the number of times IP a exists in these sets: $n_a = \sum_t \sum_{u \in U_t} 1(a \in A_{(t,u)})$. Note that for a given topic t, $\sum_{u \in U_t} 1(a \in A_{(t,u)})$ is equal to one or zero. Either the IP address a shows up in one of the $A_{u,t}$ or it does not. It cannot show up more than once because, for a single topic, usernames exhaustively and completely enumerate the space of IP addresses. However, it *can* show up zero times, depending on the number of unique usernames for a given topic.

We now consider the setting in which we consider each IP generating a post (e.g., a username in a topic) with probability π_a. Empirically, we would like to test whether $\pi_a > 0$ for all a. Moreover, we would additionally like to estimate, for a given username, the probability that a given post was generated by IP a. Our estimation procedure is confounded by the noise added by the hash procedure. We now enumerate a simple data generating process that shows how we exploit multiple postings across topics to back out estimates for π_a.

To fix ideas, let us first focus on the perspective of a given IP a. For a given topic t, we observe $k_t = |U_t|$ usernames. We consider two possible states of the world.

1. First, $y_{at} = \sum_{u \in U_t} 1(a \in A_{(t,u)}) = 0$. That is, we do not observe the IP in our collection of IPs for each username observed in the data. This means that IP a did not generate the post, which occurs with probability $(1 - \pi_a)$, **and** IP a did not occur in one of the possible $A_{u,t}$ sets generated by a different IP address (e.g., the noise). This probability

[9] Because the uniformity property holds in expectation, the exact number of matching IP addresses for any single topic-username observation can vary. The number of matching IP addresses for any of the 5,184,896 topic-username combinations observed on EJMR varies between 64,195 and 66,774 with a mean of 65,537, which is $2^{16} + 1$.

[10] Note for now, we ignore repeated posts on a thread by the same username, and assume this is a single data point. This ignores the possibility of a "collision" wherein two distinct IPs are both assigned the same username.

is significantly more complicated.

The probability of IP a not occurring in one of the possible $A_{u,t}$ set is a function of two parameters: a) the random probability of being in one of the sets, which is roughly $1/2^{16} = 1/\kappa$ thanks to the uniformity property of the algorithm and b) the number of unique usernames k_t in topic t. Specifically, this is a hypergeometric distribution with k_t draws where there are $2^{16} - 1$ balls in one urn, and 1 in the other. The probability of the IP address not being generated is then $q(k_t) = \binom{2^{16}-1}{k_t} / \binom{2^{16}}{k_t}$.

This implies that the probability we do not observe the IP a in our collection of IPs for each username u observed in the data for a given topic t is $(1 - \pi_a)q(k_t)$.

2. Second, we may observe the IP in our collection of IPs: $y_{at} = \sum_{u \in U_t} 1(a \in A_{(t,u)}) = 1$. This can happen because *either* the IP address a indeed posted or because of noise if the IP address did not post. The probability of this event is $\pi_a + (1 - \pi_a)(1 - q(k_t))$.

This helps us understand the challenge of identifying whether a post is done by a certain IP. For any single topic, we are unable to distinguish between the noise generated by the hashing and an IP's true propensity of posting. More usernames in a given post (i.e., higher k_t) also does not help our identification, but does increase the likelihood of being observed in the set of data. What *does* help is posting across topics because the randomness of the hash scrambles the binning into usernames and creates noise that is independent across topics.

We can now consider multiple topics m. If we observe the IP address a exactly n_a times across these topics, we can define the probability of this state of the world using a joint likelihood that treats each topic as independent. Our observed data is a set of observations, $y_{a,t}$ of whether we observe IP a in the set of possible IPs for topic t. Let the combined realized data be denoted as an $m \times 1$ vector $\boldsymbol{y}_a$ of the observed binary outcomes $y_{a,t}$, and the random variable associated with this data $\boldsymbol{Y}_a$. Then, the joint probability is

$$Pr(\boldsymbol{Y}_a = \boldsymbol{y}_a) = \prod_{t=1}^{m} \Big((1 - \pi_a) q(k_t) \Big)^{1-y_t} \Big(\pi_a + (1 - \pi_a)(1 - q(k_t)) \Big)^{y_t}. \tag{2}$$

Under the null hypothesis that $\pi_a = 0$, this expression simplifies to

$$Pr(\boldsymbol{y}_a = \boldsymbol{y}) = \prod_{t=1}^{m} \Big(q(k_t) \Big)^{1-y_t} \Big((1 - q(k_t)) \Big)^{y_t}. \tag{3}$$

If $k_t = 1$, this is identical to a binomial distribution. When k_t can vary across topics, this probability is a mixture of binomials and is also referred to as a Poisson Binomial distribution (see Tang and Tang (2023)). It can be written more succinctly in terms of n_a, the number of

times that IP a is observed across the different topics:

$$Pr(n_a = k) = \sum_{A \subset [m], |A|=k} \prod_{t \in A} p_t \prod_{t' \in A^c} (1 - p_{t'}), \tag{4}$$

where $[m]$ is the set of topic indices $\{1, \ldots, n\}$, and hence the summand is the sum over subsets A of the indices that are size k.[11]

This suggests two possible approaches for identifying IP addresses. First, we can consider hypothesis tests of whether $\pi_a > 0$, using a Poisson-binomial test, and adjusting appropriately for multiple tests. Second, we can directly estimate π_a. We denote this first approach the "algorithmic" approach and explain it in detail in the remaining part of this section.

Concretely, the algorithmic approach requires that we estimate the p-values for each a in a set $A_{(t,u)}$ and for each (t, u) assign the IP address with the lowest p-value if that p-value is below a threshold p^*. This assignment procedure suffers from two potential issues.

First, the above analysis involves a classic multiple-hypothesis testing problem (Hochberg and Tamhane, 1987; Benjamini and Hochberg, 1995) because for each post we test whether any of the approximately 65k IP addresses from which it could have originated rises to the required level of significance. Since we are interested in avoiding false positives (e.g., we want to control the overall size of our family of statistical tests), we want to choose a p^* that is sufficiently conservative. We describe the method we chose in the following section.

Second, the above discussion considered the universe of all (t, u) combinations in our data when examining the distribution of n_a. However, this can lead to many false positive assignments, since a given IP will show up $N(1/\kappa)$ times randomly due to the hashing function, even under the null. If N is large, then low p-value IP addresses may show up randomly for posts that were posted by IPs that show up infrequently (and hence have higher p-values). To solve this issue, we window the data in the different time intervals. Using this approach, we are able to reduce the expected number of incorrect assignments to zero.

2.4 Assignment of Posts to IP Addresses

The foregoing formalization endows us with the ability to say which IPv4 addresses are "active" on EJMR in some window of time. However, it does not say precisely how one would assign an observed post to a *particular* IP address. A proper hierarchical Bayesian model for doing so would likely describe humans, some of whom are economists, probabilistically acquiring and releasing IP addresses, viewing EJMR topics, and selecting in which topics to post according to their individual preferences. Such a model is likely under-specified and, for the moment, beyond our abilities. In its stead, we present a model that is *practical* in the sense of being

[11] When $p_t = p$, this simplifies to $\binom{m}{k} p^k (1-p)^{(m-k)}$, the binomial distribution.

both intuitive and tractable.

The intuition for our practical model is as follows. Consider a post on EJMR. Like all other posts, we know from our enumeration procedure that this post has about 65k IP addresses from which it might have originated. These IP addresses can "explain" the post's username. But, imagine that one of these IP addresses *also* explains twenty other posts made around the same time. What is the likely origin of the post? It is, we contend, this highly explanatory IP address.

Our reasoning relies on the following sparsity argument. Not many humans study economics, fewer still of those have PhDs, post on EJMR, and are active in a given week. Furthermore, many lines of research show that contributions to online media are highly concentrated: 20% of users often produce 80% of the content (Guo et al., 2009). With these intuitions, we turn the IP address assignment problem into an optimization problem. We adopt a simple rule by which we choose to assign posts to the smallest set of IP addresses that can explain them in a given window of time. However, we do not assign *all* posts to IP addresses. We use the foregoing statistical model to limit our candidate IP addresses to those which are above some threshold of apparent activity that is improbable by chance. The structure of our data allows us to determine this threshold in a manner that minimizes incorrect attributions. In other words, rather than saying "this post came from this IP with probability X" we are saying "this post came from this IP if you use this sparsity rule" and we describe some of the error properties of that rule.

Our assignment procedure is structured as follows. First, we order all the posts by time and bin them by GMT date. We consider all the posts on a single day. From this day we extend a window of time three days into the past and three into the future, thereby collecting a week's worth of posts. For each of the posts in this week, we gather the explanatory candidate IPv4 addresses which we obtain from the hash inversion procedure described before. For each of the roughly 4.3 billion IPv4 addresses, we count the number of unique topic-username (t, u) pairs that the IP potentially explains in the week.[12] Then we compute the p-value for this count for each IP address. Recall from the previous section that these counts follow a Poisson binomial distribution. This distribution has an intractable normalization constant for data of our size, which is the primary reason why we process posts day-by-day. We approximate the Poisson binomial probability mass function for the counts just once per day using a fast Fourier transformation (Biscarri et al., 2018).

For each post on the target day, we assign the post to the IP address with the lowest p-value, but *only* if that p-value is below some threshold p^*, which is determined in a manner we describe shortly. Having attempted to assign all the posts on the target day, we move to

[12]Recall that given the nature of the username allocation algorithm, an IP address posting multiple times in the same topic is assigned the same username.

the next day and repeat the procedure.

To recover IP addresses that do not post as frequently during our relatively short 7-day window but still post regularly over a longer time periods our procedure considers two additional time windows of 31 and 91 days. Users may have different patterns of posting and the different window sizes allow us to discover these different users.

How should we determine p^*? In other words, how can we know when an IP address is overrepresented in a particular window of time and that its observed explanatory power does not arise by chance in the noise component of equation (3)? Fortunately, we have a very accurate way of modeling the noise distribution by "shifting" the characters from which the EJMR username is drawn in the SHA-1 hash. Instead of using positions 10-13 of the hash, we use positions 11-14 and ask the following question. What if the IP addresses that could have generated the username for this post were not those with a SHA-1 hash containing the username at position 10, but rather at position 11? Both sets of IPs are roughly 65k in size. The position-10 set is guaranteed to contain the true IP address and all the other IP addresses contained in it are "noise." In contrast, the position-11 set contains *only* noise. The true IP address might be therein, but if so, only by chance.

We use this insight to determine p^*. We repeat the entire hash inversion as if the EJMR username was drawn from positions 11-14 and then repeat the post-IP assignment procedure described above. Then, we calculate the p-value thresholds p^* such that we would obtain *zero* assignments of posts to an IP address for position-11. That is to say, using the (incorrect) position-11 hashing set and these thresholds, *none* of the roughly 7 million posts observed on EJMR would be assigned an IP address. The p-value thresholds we found are 1.37×10^{-10} for the 7-day window, 2.51×10^{-11} for the 31-day window, and 1.39×10^{-11} for the 91-day window. We then use these same p^* thresholds for the correct hash positions (position-9 prior to July 13, 2013 and position-10 thereafter until May 17, 2023).

The p-value thresholds above are clearly very small. This is because our method of determining p^* naturally adjusts for multiple hypothesis testing. For each window (7, 31, and 91 days) we are conducting approximately 7 million $\times$ 65k $\approx$ half a trillion hypothesis tests. Thus, a p-value threshold with a low *overall* error rate will be quite small. The number of IP addresses that never posted to EJMR but that we mistakenly assign to a post is, in expectation, *less than one*. However, there are potentially other varieties of error, which we discuss later in this section.

In the end, we completed the assignment procedure using three different starting positions of the hash (9, 10, and 11). Figure 3 shows the average minimum p-value of IP addresses for all the posts in a given week for each of these hash positions over time. The value at each point in this graph is most clearly described by the two-step procedure we use to calculate

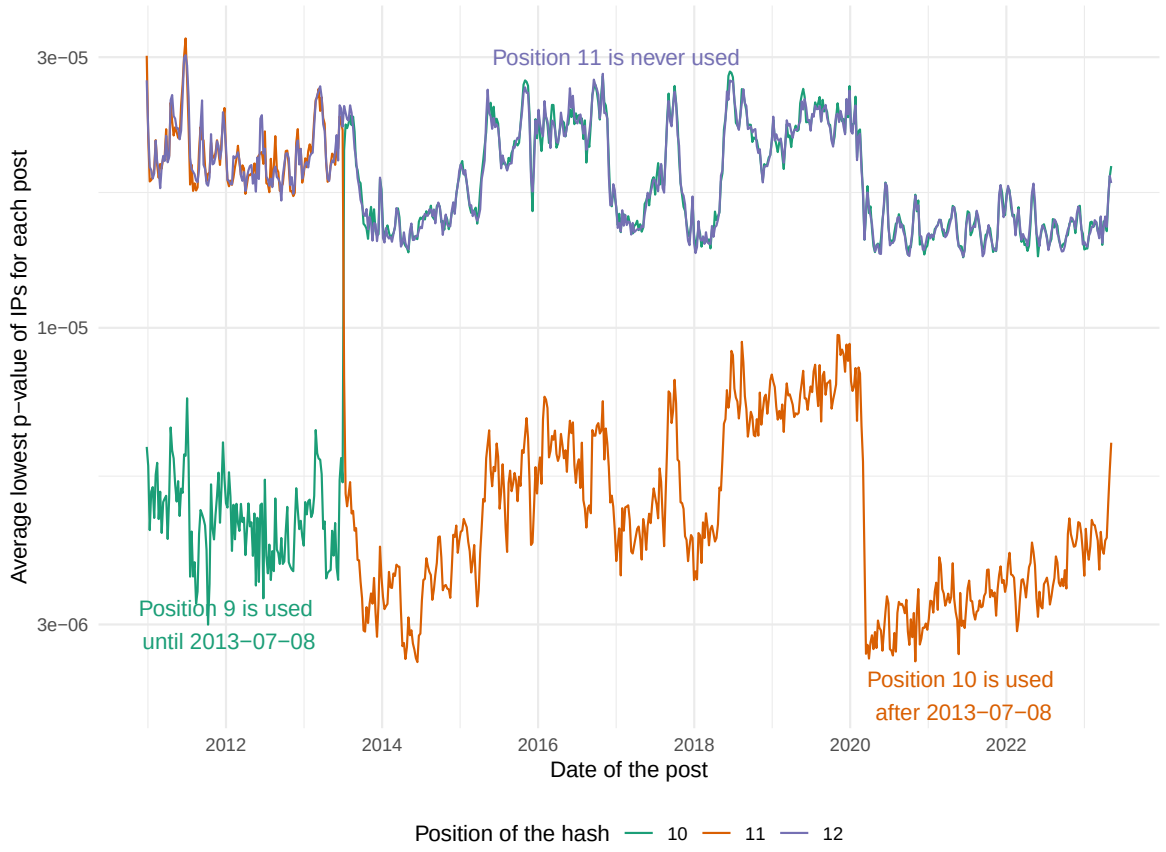


Figure 3: Average minimum p-value of posts in a given week for different hash positions over time. We employ the following two-step procedure. First, we find the IP address that has the lowest p-value for a given post and refer to its p-value as the minimum p-value of a post. Second, we calculate the mean of the minimum p-value of a post across all posts in a given week. The graph clearly shows what hash position was used for the EJMR username at each date. The orange line showing the position-10 p-values is toward the bottom of the graph because EJMR usernames started at position-10 for most of the website's history. On July 8, 2013, the site administrator likely made a database error that "shifted" the usernames one position left. For this reason, the position-9 p-values are lower before this date. When a position is not "in use," its p-values closely track the position-11 "noise" distribution. Note that none of the posts to which we actually assign IP addresses would be visible on this graph as our p^* thresholds are on the order of 10^{-11}. These low p-values are in effect what is pulling down the green and orange lines which are weekly averages.

it. First, we find the IP address that has the lowest p-value for a given post and henceforth refer to this p-value as the minimum p-value of a post. Second, we calculate the mean of the minimum p-value of a post across all posts in a given week. The graph clearly shows what hash position was used for the EJMR username at each date. The orange line showing the position-10 p-values is toward the bottom of the graph because EJMR usernames started at position-10 for most of the website's history. On July 8, 2013, the site administrator likely made a database error that "shifted" the usernames one position left. For this reason, the

position-9 p-values are lower before this date. When a position is not "in use," its p-values closely track the position-11 "noise" distribution. Importantly, none of the posts to which we actually assign IP addresses would be visible on this graph as our p^* thresholds are much smaller and on the order of 10^{-11}. These low p-values of posts to which we assign IP addresses, are in effect what is lowering the green and orange lines which are averages across all posts in a week.

Having confirmed the cutoff date between position-9 and position-10, we elected to use only the position-9 assignments prior to July 8, 2013 and only the position-10 assignments afterward.[13] In total, we assigned IP addresses to 4,692,946 of the 6,912,773 EJMR posts for which we have both a topic ID and username, or 66.1% of posts over the period spanning December 21, 2010 to May 10, 2023. These posts originate from just 47,630 distinct IP addresses.[14] Most of our assignments come from the 7-day window procedure. Roughly speaking, if an IP address was the source of posts on more than about a dozen topics in a week, that IP address is identified by our method.

Figure 4 plots the cumulative distribution functions of the minimum p-values of the posts for different hash positions. The orange line plots the CDF of the minimum p-values for posts with the incorrect position-11 hash which only contains noise. Comparing this line with the approximate p-value $p^* \approx 10^{-11}$ which is represented by the dashed vertical line, it is evident that not a single post of the 6,912,773 EJMR posts for which we have both a topic ID and username would be assigned to an IP address because no post would have a sufficiently low minimum p-value. In contrast, the green line shows the CDF under the correct hash (position-9 before July 8, 2013 and position-10 afterward). 67.9% of posts have minimum p-values below p^* and for almost 30% of posts the minimum p-values are smaller than 10^{-50}.

Earlier, we claimed our method was unlikely to assign posts to IP addresses that are truly inactive on EJMR. We identified a convenient way of testing that claim because the IPv4 address space contains certain "reserved use" IP addresses from which no traffic should "legitimately appear on the public internet" (Cotton et al., 2010). These are the so-called *bogon* addresses, which are nearly 600m in number and thus occupy 13.8% of the entire IPv4 address space. We know that any assignments we made to bogon IP addresses were surely in error. However, out of the 4,692,946 posts to which we assigned IP addresses, the number of posts our procedure assigned to bogons is *zero*.

There is, however one type of significant—but estimable—error in our assignments. This error arises from high-posting IP addresses "stealing" posts from the *true* posting IP address. Of course, we do not observe the true posting IP address. To gain an intuition for this

[13] On the cut-off date we allow either hash position.

[14] We also have 185,338 posts that have either no topic ID or no username. These cannot be assigned to IP addresses.

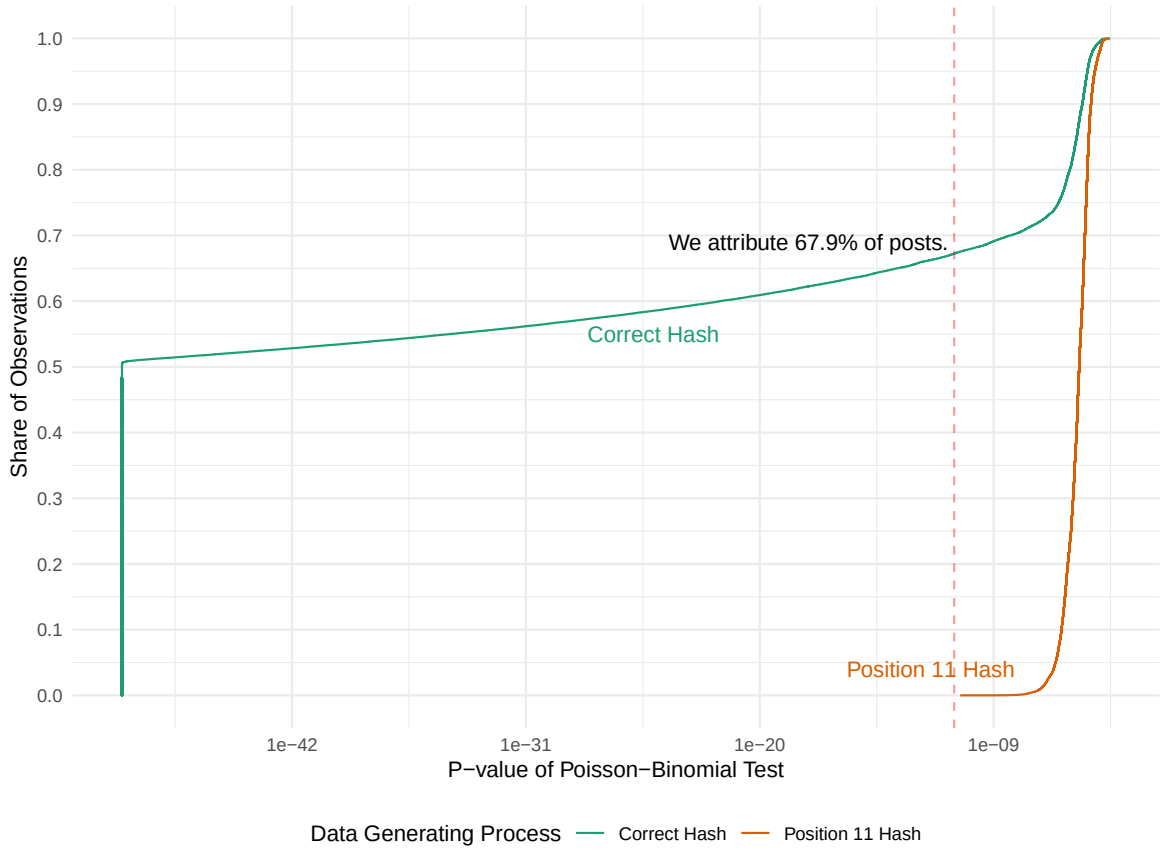


Figure 4: Cumulative distribution functions of the minimum p-value of posts for different hash positions. The orange line plots the CDF of the minimum p-values for posts calculated based on the incorrect position-11 hash. The green line shows the CDF under the correct hash (position-9 before July 8, 2013 and position-10 afterward). The approximate p-value threshold $p^* \approx 10^{-11}$ is represented by the dashed vertical line. For position-11, none of the 6,912,773 EJMR posts for which we have both a topic ID and username would be assigned to an IP address because no IP has a sufficiently low minimum p-value. In contrast, 67.9% of these posts have minimum p-values below p^* and for more than 50% of posts the minimum p-values are smaller than 10^{-50}.

type of error, imagine the following situation. A one-time EJMR user posts to some topic and receives username ab34. However, it just so happens that a highly active IP address would also receive that same username if it posted in that topic. That is, an assignment on this highly active IP address occurs by chance in what we have been calling the "noise" component of the SHA-1-based username. In our assignment scheme, we would mistakenly assign the post to the highly active IP. Recall that our scheme is basically an optimization that assigns posts to the smallest set of IP addresses that explains the posts subject to a significance threshold. We do not have as precise a model as we would like for how often such a misassignment occurs. However, we have a rough estimate. First, note that this situation is fairly rare. Our event windows are small and the number of highly active IPs at any given

time is very small relative to the total number of IPv4 address. Second, we believe that this kind of error is maximized when the window size used for assignment is maximized and when the IP in question is most highly active. That is, highly active IP addresses can "steal" the most posts and the opportunity to "steal" is largest when the window size is largest. Consider the maximal window size, a window spanning all 13 years of EJMR activity for which we have data. The maximally active IP address has about 47k posts on EJMR. That IP address would be "explanatory" by chance for about one in every 65k topic/username pairs, or about 80 pairs over the 5,184,896 observed in our data set. That would mean about 106 posts (based on the ratio of posts to unique topic/username pairs) out of the 47k should be expected to have been assigned in error or about 0.2%. Of course, most of our assignments happen for the 7-day window size. For less active IPs, there is substantially less opportunity to "steal" assignments. As a result, we expect this error rate to be low.

2.5 Linguistic Analysis

This section contains offensive speech that some readers may wish to avoid. Those readers should skip to Section 2.6. We analyzed the linguistic character of EJMR posts and topics using a variety of machine learning techniques. However, due to the extensive use of obfuscation on the site, many posts required some pre-processing. For example, consider the following posts:

- "Given women get free spots, blks and latins get free spots, it basically means you need to be far far right tail if u are a yt or azn homegrown American." (2022-12-27)
- "Mold-fa//g//g//ot, I will split your a//s/s in two with my HUMONGOUS super HARD shalong. You will be squealing like the little beia/tch that you are." (2020-01-28)
- "those d4mn j3ws had no morals either." (2022-08-13)
- "Hey a$$h01e, I left you a message earlier too. I will be there in Boston to FIEK and RAEP you, so cover your $hitty a$$ and your mouth now." (2014-12-26)

These posts are obfuscated to such an extent that we found most machine learning models failed to accurately classify them as toxic. To address this, we developed software to deobfuscate such speech. First, we classified posts into commonly occurring natural languages on EJMR (Stahl, 2023): English, German, Chinese, Korean, and a few others. Then we collected high-frequency non-English words in the English posts which we used to develop a dictionary mapping text like "f**k," "secks," and "GTFO" to canonical forms. We used this dictionary to deobfuscate some of the most commonly obfuscated terms.

Then, we checked each word in each post for common symbol-based obfuscations like "fa//g//g//ot," removing symbols where doing so resulted in an English word or well-known profanity. Finally, we transformed so-called leetspeak—such as "d4mn j3ws"—to its canonical form. We did this by attempting common leetspeak substitutions and checking if those substitutions resulted in an English word or a well-known profanity. Our goal in this effort was not perfection, but rather *some* improvement in the performance of machine learning models for this content. Numerous obfuscations remained after our triage. For example, the use of "yt" is context dependent: sometimes it means "white" and other times it means "YouTube." So we chose not to transform certain out-of-corpus words like "yt" when we found them.

Having deobfuscated all posts, we ran each post through a number of transformer-based machine learning models. We selected models that are state-of-the art and that, based on our informal inspection, also appeared to perform well on EJMR content. For sentiment detection, we selected the default Huggingface sentiment model. This is a checkpoint of DistilBERT (Sanh et al., 2019) fine-tuned for sentiment detection (Hugging Face, 2023) on the Stanford Sentiment Treebank (Socher et al., 2013). We selected a similar fine-tuned BERT model for detecting misogyny (Attanasio et al., 2022). For toxicity we selected ToxiGen Roberta (Hartvigsen et al., 2022). This is a checkpoint of the Roberta model (Liu et al., 2019) fined-tuned for toxicity detection.[15] We also re-created all the word-count measures used in Wu (2020).

2.6 Comparison with Reddit

Reddit is a social media platform where users can post content, comment, and vote on posts. Much of that is done pseudonymously. As of 2023, it is the 10th most-visited website in the world. It is organized into "subreddits," which are topic-specific communities. Subreddit moderation on Reddit is carried out by administrators who are either official Reddit employees or individuals selected by specific community members. Reddit bestows a degree of autonomy upon these subreddit moderators, enabling them to determine the permissible and unacceptable content within their respective subreddits, as long as they remain within the bounds of site-wide rules. This notable flexibility has paved the way for the emergence of a wide array of subreddits, some of which have stirred up controversy. The decentralized nature of Reddit's moderation, coupled with user anonymity and the absence of robust fact-checking mechanisms, has rendered the platform susceptible to the dissemination of misinformation and the reinforcement of echo chambers, ultimately fostering distorted worldviews among its

[15] A nearly identical Toxigen-based model was used to measure toxicity in Meta's recent Llama-2 large language model (Touvron et al., 2023).

user base (Cinelli et al., 2021).

We retrieved a list of the 1,500 subreddits with the most subscribers from https://www.reddit.com/best/communities. From those we selected the top 1,000 subreddits that had at least 10,000 posts in the Pushshift Reddit Dataset Baumgartner et al. (2020), which contains roughly all Reddit content from the site's inception through 2018. Clearly, some popular subreddits in 2023 did not exist in 2018 and some subreddits that were popular in 2018 were likely not popular in 2023. From each of these 1,000 most popular subreddits, we downloaded all "utterances"—these are analogous to "posts" on EJMR. For each subreddit, we randomly, uniformly sampled 10,000 posts. We subjected these posts to the same linguistic analyses that we used for EJMR posts. They were deobfuscated and classified using the same models as EJMR posts.

2.7 Geolocation of IP Addresses

We use the commercial IP2Location database to obtain country, city, latitude-longitude, zip code, ISP, and domain information for all the IP addresses we identify.

Internet users have multiple IP addresses across the devices and the networks they use. Recent research suggests that the average consumer IP address in the United States is held for about 19 days and about 87% of internet users will have, at any time, at least one IP address that is used for more than a month (Mishra et al., 2020). Because universities tend to have generous IP blocks—particularly elite universities—it seems likely that IP retention in these institutions will be longer. In addition, geolocation for university IPs is particularly accurate (Saxon and Feamster, 2022).

2.8 Time Stamps

Unfortunately, EJMR does not display exact post times. Depending on the age of the post, it only displays whether it occurred "m/h/d/m/y minutes/hours/days/months/years ago." This makes it, a priori, difficult to assign exact time stamps to posts, especially to older posts. However, EJMR provides two additional pieces of information. First, it has an RSS feed which displays the most recent 10 posts along with the exact time stamps (year, month, day, hour, minute, second) in every topic. Second, every post on EJMR has a unique, auto-incrementing integer post ID starting at 1 on December 17, 2010. This post ID increases one-by-one for all the posts across all the topics on the site.

We downloaded the RSS feed and the Wayback Machine for all the 642,247 topics. This gave us exact time stamps for a total of 3,689,727 posts. For the remaining 3,408,384 posts, we assigned the time stamp based on the auto-incrementing post ID by linearly interpolating between any two posts with known exact time stamps. Because the posts with exact time

stamps are very evenly distributed we were able to accurately assign time stamps for all posts without exact time spots. The average time difference between posts with exact time stamps that have some posts without exact time stamps in between them is only about 3 minutes and even at the 95th and 99th percentile this difference is smaller than 10 minutes and 23 minutes, respectively.

3 Results

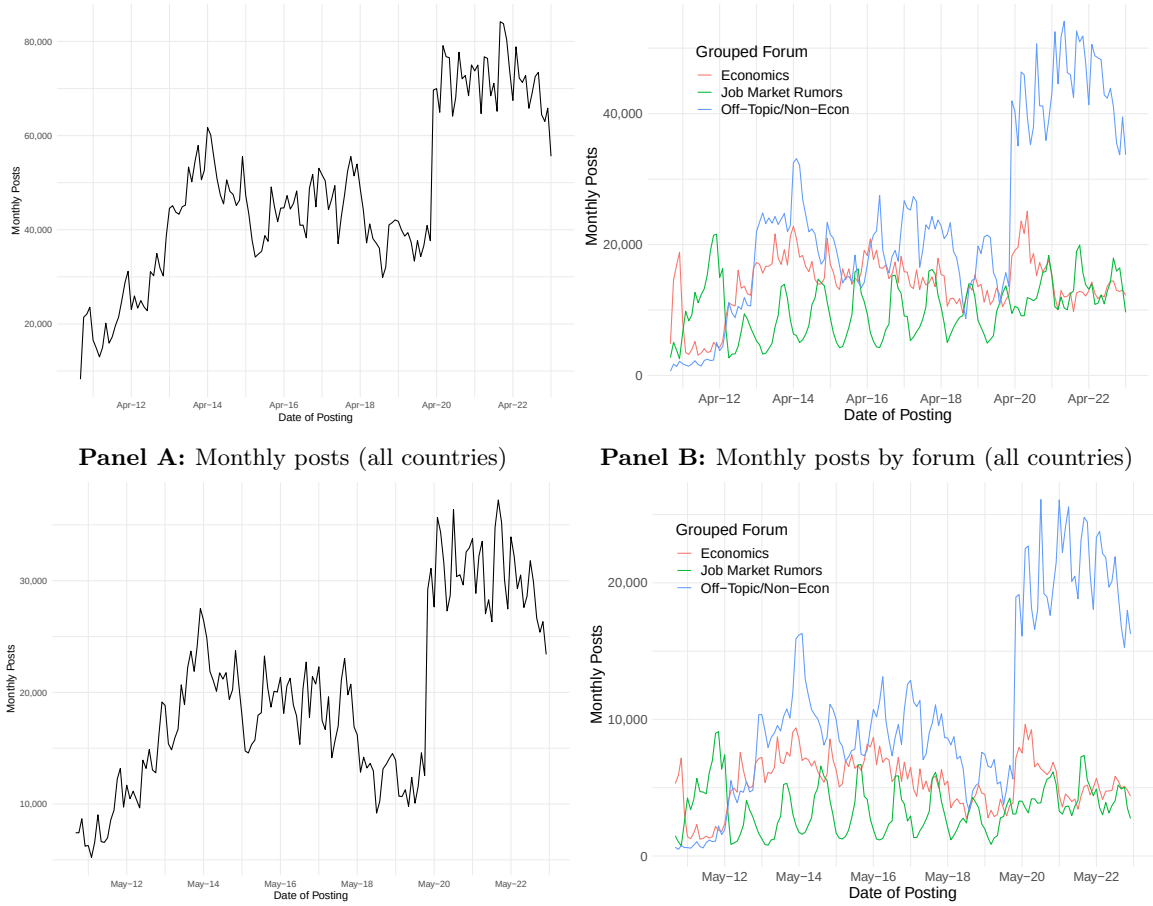
3.1 Descriptive Statistics

We obtained EJMR content from both http://www.econjobrumors.com and http://archive.org. In total, our data included 7,098,111 posts from 695,364 topics on EJMR between December 17, 2010 and May 10, 2023. There are 185,338 posts for which we do not have topic identifiers or usernames mostly because they originated from registered users for which the site does not display hexadecimal usernames. With our methods these posts are, by construction, unassignable to IP addresses. The remaining 6,912,773 posts for which we have topic identifiers and usernames, are assignable posts which, in principle, can be assigned to IP addresses. From these assignable posts we recover 47,630 distinct IP addresses.

3.2 Time Patterns

Panel A of Figure 5 shows that the posting frequency on EJMR steadily rises between December 2010 and April 2014 and then remains relatively stable at around 40,000 monthly posts between 2013 and the beginning of 2020. However, the posting intensity jumps to around 70,000 posts per month in March 2020 with the beginning of the COVID-19 pandemic and, until recently, has remained at this elevated level. Panel B of Figure 5 makes the cyclicality of the academic job market apparent. The green line which plots the monthly posts in the Job Market Rumors forums always peaks in December and January during the busiest part of the job market for academic economists.

The increase in EJMR posting frequency induced by the COVID-19 pandemic appears to be driven by two factors. First, as can be seen in Figure 5, the aggregate posting increase comes entirely from a sharp increase in the number of posts in the Off-Topic/Non-Econ forums which quadruples in size. There is also short transient increase in the number of posts in the Economics forums, but this increase subsides relatively quickly after a year. Second, as can be seen in panels C and D of Figure 5, the increase is primarily driven by IP addresses located in the United States whose monthly posting volume tripled from around 10,000 to over 30,000 posts per week and remained high.



Panel A: Monthly posts (all countries)

Panel B: Monthly posts by forum (all countries)

Panel C: Monthly posts (United States)

Panel D: Monthly posts by forum (United States)

Figure 5: Number of monthly EJMR posts (left panels) and by grouped forums (right panels) over time for all countries (top panels) and the United States (lower panels). The left panel of the figure shows the number of monthly posts from December 2010 to April 2023. There is a marked increase in posting activity that coincides with the start of the COVID-19 pandemic in Europe and the United States in March 2020. The right panel shows the number of monthly posts in three groups of forums. Economics contains all forums with general economics discussions. Job Market Rumors contains all forums related to the academic job market including both junior and senior hiring. Off-Topic/Non-Econ contains all other forums.

The COVID-19-induced increase in posting activity on EJMR however is not entirely confined to the United States. Other countries from which a large number of postings originate such as Canada, the United Kingdom, Italy, France, and Germany also experienced large increases in 2020, mostly in the Off-Topic category, as can be seen in Figure 6. However, unlike the United States the posting intensity in these countries mostly returned to pre-pandemic levels. Posts for which we cannot assign an IP address (bottom middle panel) display a relatively steady increase across all three forum groups over time and also have very strong posting cyclicality in the Job Market Rumors category.

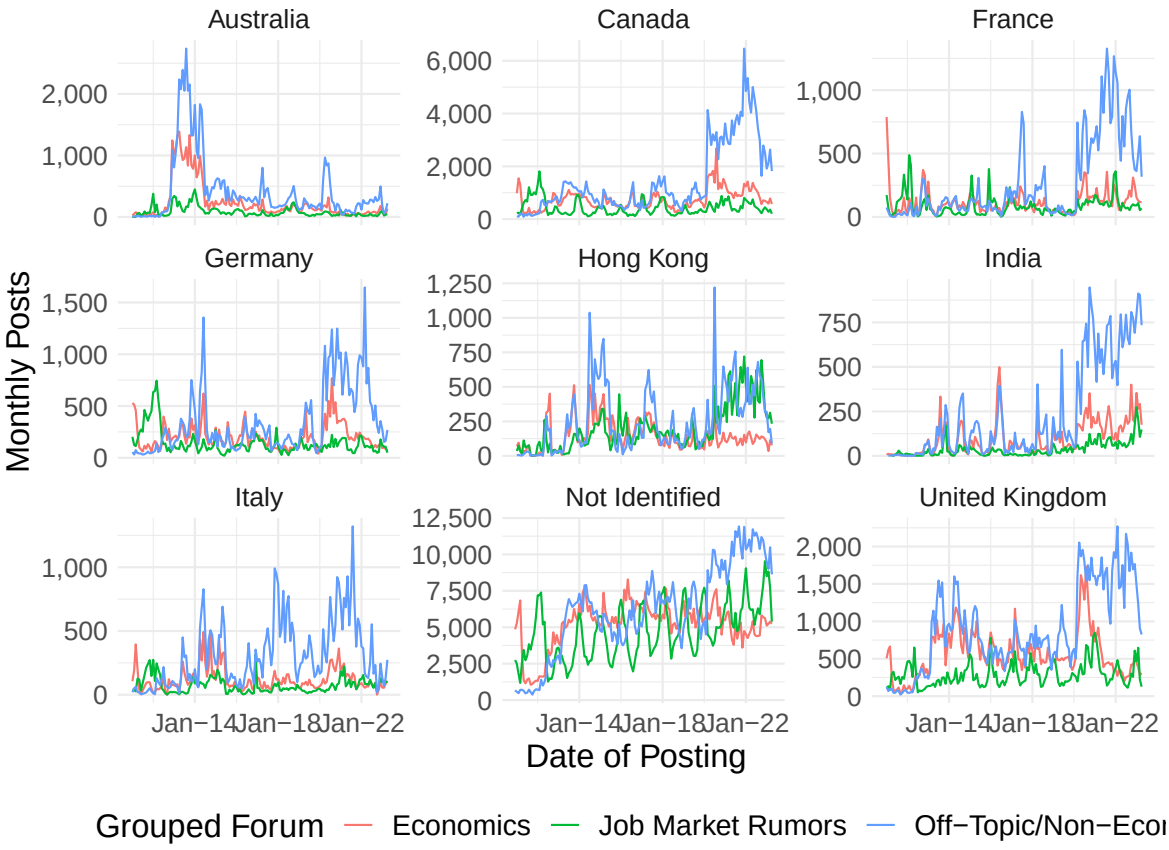


Figure 6: Monthly posts by grouped forums and by country (excluding US). The figure shows the distribution of monthly posts across the three large forum groups for the eight countries with the largest number of posts after the United States and for those posts for which we do not assign IP addresses.

EJMR users tend to primarily post during US work hours and to a lesser extent in the evening. Figure 7 shows total number of posts per minute by year from 2011 to 2022. The graph reveals that usage overall increased since the onset of the COVID-19 pandemic, but it did not change the overall pattern of EJMR users primarily posting during work hours.

This pattern becomes even more apparent when using the country location of the posting IP addresses. Figure 8 reports the distribution of posts across the time of day for the six countries with the largest number of posts: Australia, Canada, Germany, United Kingdom, Hong Kong, United States. Adjusted for their respective time zones EJMR users tend to post in the afternoon and in the evening, but less so during the morning or at night. However, as noted previously, the majority of the posts originate from IP addresses located in the United States.

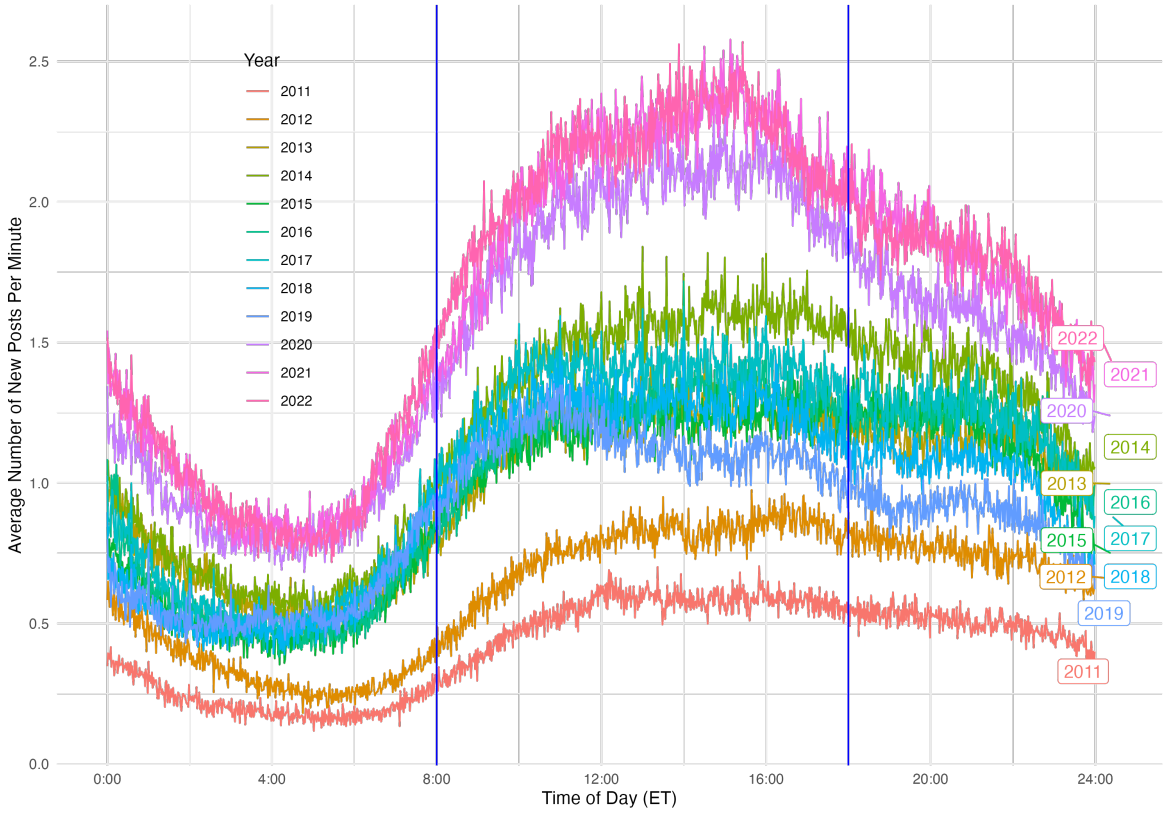


Figure 7: Distribution of posts across time of day (US Eastern Time). The figure shows the average number of EJMR posts per minute for a given year. The vertical blue lines show standard work hours for US Eastern time.

3.3 Geographical Distribution

The majority of the assigned EJMR posts originate from IP addresses located in the United States. As noted in Section 2.2, with our methodology we assign 67.9% of assignable posts to IP addresses.

Figure 9 shows that we are able to assign 67.8% of assignable posts to particular countries using IP2Location. Among posts with geolocated IP addresses, 61.9% originate from the United States with Canada (8.3%) and the United Kingdom (5.5%) a distant second and third. The rest of the posts with geolocated IP addresses come from other countries with significant research institutions in economics and finance such as Australia, Germany, Hong Kong, Italy, and France. There is also a substantial share of geolocated posts (13.6%) from other countries in the rest of the world.

An additional sanity check for the accuracy of our IP assignment and subsequent geolocation is whether the language that EJMR posters use corresponds to the country of origin of their IP address. Using the language classification of Stahl (2023) we show in Table 1 that

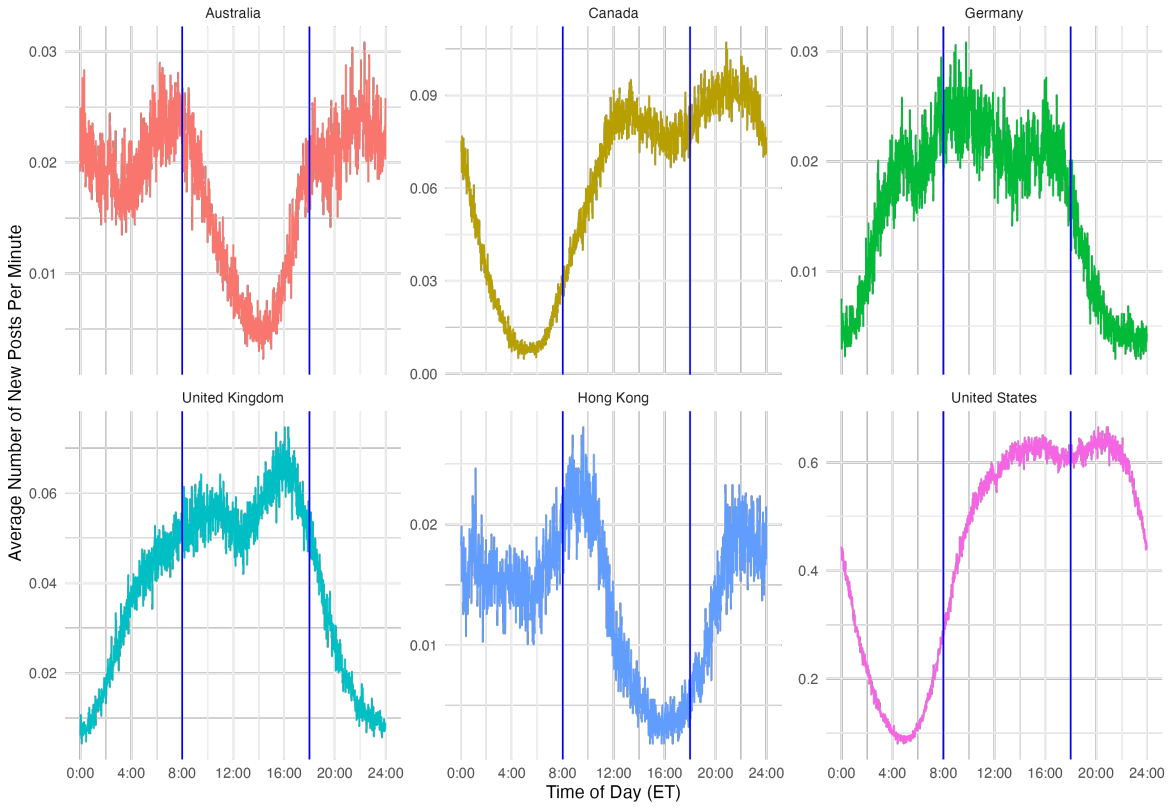


Figure 8: Distribution of posts across time of day (US Eastern Time) by country. The figure shows the average number of EJMR posts per minute for Australia, Canada, Germany, United Kingdom, Hong Kong, and the United States. The vertical blue lines show standard work hours for US Eastern time.

the dominant non-English language in all major non-English speaking countries in our data is indeed the country's native language. This pattern is particularly pronounced for Brazil, China, Germany, Hong Kong, and Korea and to a lesser extent for Spain, Portugal, and the Russia.

Beyond country of origin the IP addresses we recover also provide much more granular information about the exact location and internet service provider of EJMR posters. Figure 10 reports the cities with the largest number of posts. There is substantial heterogeneity driving the ranking of these cities. By far the largest number of posts originate from Chicago. However, the number of unique IP addresses from which posts attributed to Chicago originate is comparatively small (879 unique IP addresses).[16] In contrast, the next two cities on the list, Hong Kong and New York City, have much fewer posts but these posts originate from a

[16]A large number of posts originate from a few IP addresses located in Naperville, IL. Based on several cross-checks with other geolocation databases, these IP addresses are misclassified and are actually located in Chicago. We correct these locations in the IP2Location data.

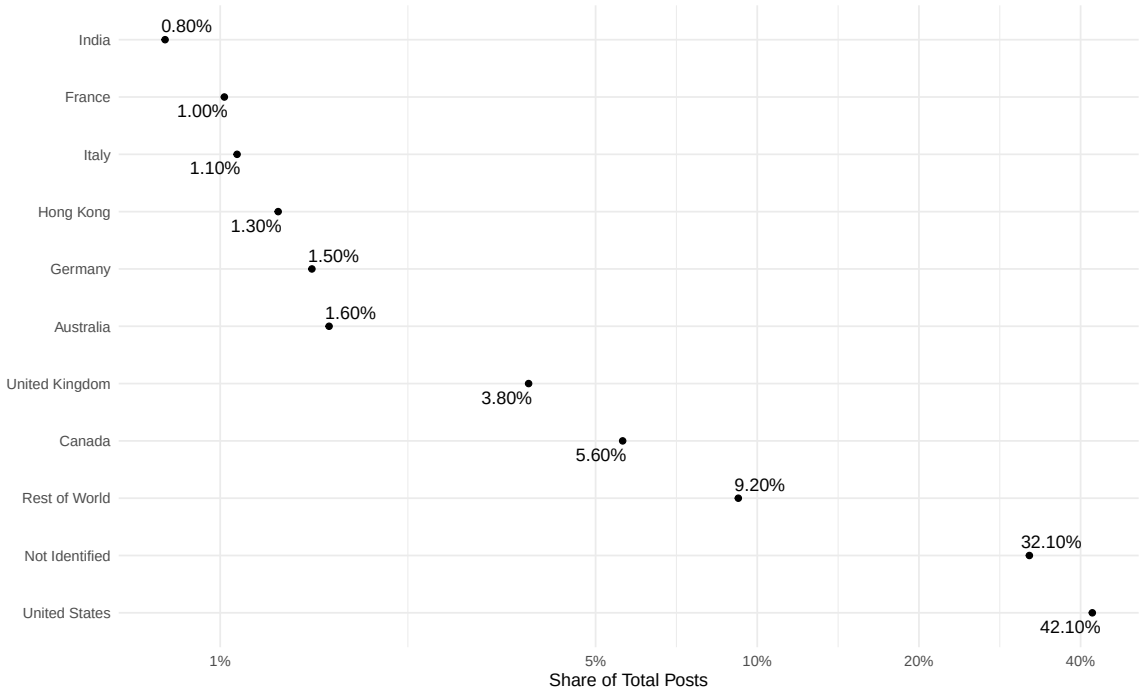


Figure 9: Distribution of posts across countries. The figure shows the share of all posts which can be assigned to a particular country. Posts for which we do not assign an IP address are in the Not Identified category.

larger set of IP addresses (952 and 1,324 unique IP addresses, respectively). Cambridge, the location of two of the leading economics departments in the world, is also among the top 5 cities and also has a smaller number of IP addresses (499 unique IP addresses) from which its posts originate. As expected from our country-level analysis, cities in the United States, particularly those with leading universities such as Cambridge or Berkeley are towards the top of the ranking despite their relatively small population size.

The association of an IP address with a contributor is not necessarily persistent, both in the short and long run. Posters may change IP addresses because of new IP assignment by their internet service provider, the use of a different device, or various other reasons. Nonetheless, IP address do persist for a significant period of time for power users of the site. In Figure 11, we plot the posting frequency for a select number of IP addresses (and their respective locations) from which many EJMR posts originate. Posts from these IP addresses end abruptly when the users are assigned new IP addresses, but all of them make a large number of posts over an extended period of time (i.e., several years).

Posting on (and not just reading) EJMR appears to be pervasive and widespread, even from devices directly connected to university networks. 10.2% of all posts to which we assign IP addresses originate directly from IP addresses associated with universities or research

	Language of Post					
Origin Country of Post	German	Chinese	Spanish	Portuguese	Russian	Korean
Germany	0.8	0.04	0.05	0.03	0.07	0.01
China	0.07	0.85	0.05	0.03	0	0
Hong Kong	0.09	0.86	0.03	0.02	0	0
Spain	0.29	0.03	0.35	0.33	0.01	0
Portugal	0.28	0.03	0.29	0.4	0	0
Brazil	0.09	0	0.12	0.77	0.01	0
Russia	0.29	0.06	0.05	0.03	0.57	0
Korea	0.2	0.04	0.06	0.07	0.01	0.62
Rest of World	0.32	0.31	0.18	0.12	0.04	0.02

Table 1: This figure shows the share of non-English posts for each country that are in the languages indicated in the six columns. These are the non-English languages with at least 1,000 posts on EJMR. Each country's primary language is in bold font.

institutions. Although some universities also are the internet service provider for some of their faculty and students (e.g., through university-provided faculty or student housing), this means that a substantial number of posts on EJMR occur while users are connected their workplace. Perhaps even more surprisingly, there are EJMR posts from identified IP addresses located at *every* leading university in the United States.

Figure 12 reports how many posts come from each of the top 25 US universities (US universities with an economics department listed in the U.S. News top 25) as a share of all the posts originating from IP addresses associated with universities or research institutions. The figure highlights the very large share that these top 25 US universities have across posts from all universities around the world as they account for more than 20% of posts from over 500 universities around the world that show up in our data.

The representation of top US universities among EJMR posters is also apparent in Figure 13 which reports the share of posts accounted for by a US university or research institution among all posts originating from IP addresses associated with universities or research institutions. Eight of the 19 institutions shown in the figure are universities ranked among the top 25 economics departments and among the top four universities contributing to EJMR, three (Stanford, Columbia, and University of Chicago) are ranked in the top 10. Among these institutions with the largest shares of EJMR posts there is also one organization that is not a university. The Federal Reserve Board employs over 400 PhD economists, many more than any single university. Given the sheer number of economists and that some of these economists contribute posts to EJMR while connected to their employer's network, it is perhaps not particularly surprising that the Federal Reserve Board appears in this list.

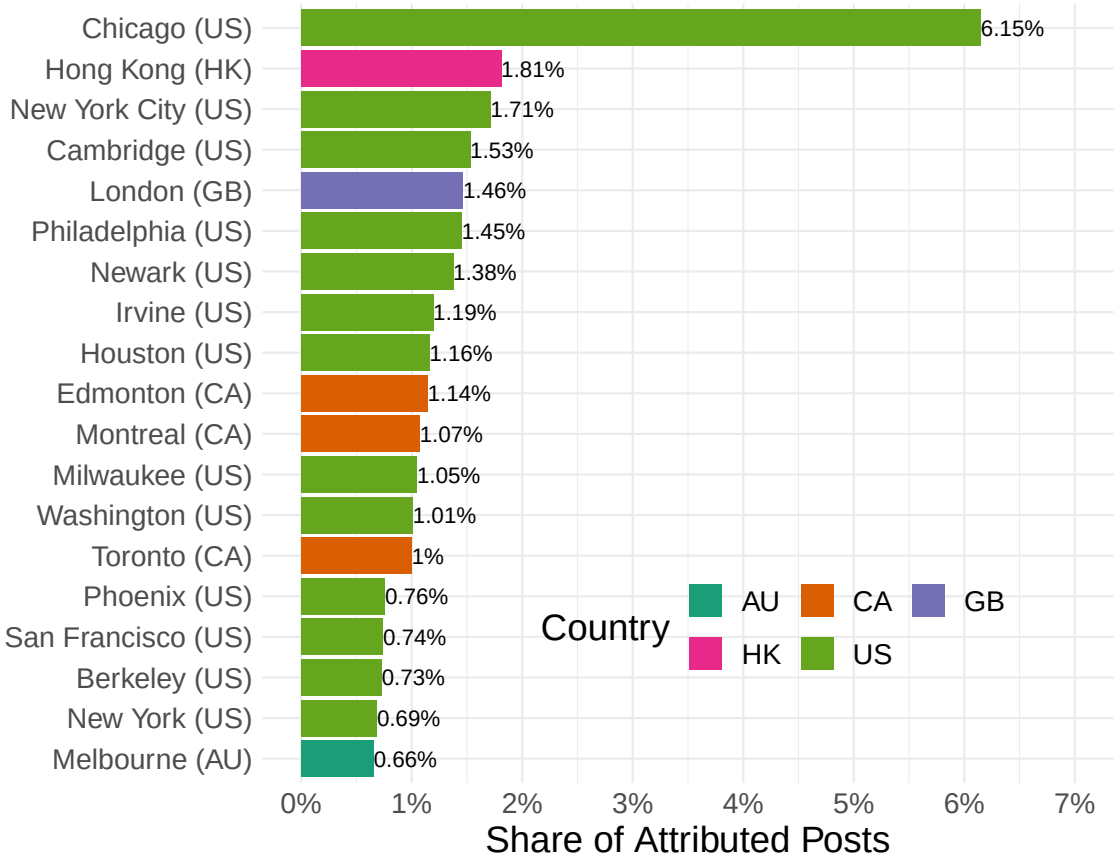


Figure 10: Share of posts with assigned IP address across cities. This figure shows the share of posts with an assigned IP address that originate from a given city. The share of posts from cities located in the United States, Hong Kong, the United Kingdom, Canada, and Australia are marked in light green, pink, purple, orange, and dark green, respectively.

The vast majority of EJMR posts originate from non-university IP addresses. Figure 14 plots the number of posts for each IP address, ranked according to the number of posts they have made. Green dots indicate that the IP address is a university or research institution. The posts-rank relationship is drawn for a log-log scale and is close to linear, providing suggestive evidence of a power law (Clauset et al., 2009). However, as we describe in detail later, the curve slightly "bulges out," suggesting a fatter tail at the top and thinner tail at the bottom of the distribution than what would be implied under a power law. Among the top 100 IP addresses by number of posts, only 15 are served by university ISPs (and none in the top 10) suggesting that power users are much more likely to be posting from residential IP addresses. However, a substantially larger proportion of university IP addresses appear between rank 101 and 1,000. Among these 900 IP addresses that post very frequently (but not quite as frequently as the top 100 IP addresses) a total of 240 are located at universities.

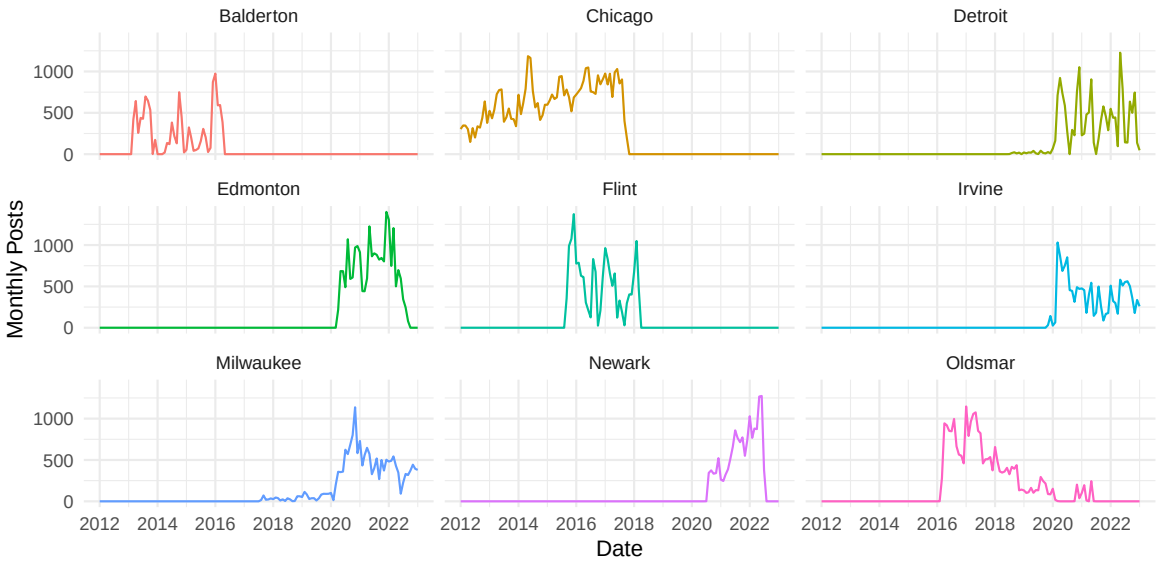


Figure 11: Monthly posts for selected EJMR power users. The figure shows the number of monthly posts over time for a select number of IP addresses from which a large number of posts originate.

Another indication that EJMR usage is pervasive throughout all echelons of the economics profession, including faculty at elite institutions, can be found in Figure 15. The figure shows the number of weekly posts that originate from the Royal Sonesta Boston, a hotel in Cambridge, Massachusetts. This hotel serves as the location of the annual NBER Summer Institute, a three-week conference held annually in July. The NBER Summer Institute is the world's leading economics research conference and attendance is by invitation only. As is evident from the figure, EJMR posts from this hotel's IP address peak every year in July except in 2020 and 2021 when the NBER Summer Institute was only held virtually rather than in person at the Royal Sonesta Boston.

3.4 Concentration of Posters and Posts

There are 6,912,773 posts for which we have the topic and the username and from which we are able to recover 47,630 distinct IP addresses. However, these posts are far from evenly distributed across the many posters on the platform. Among the posts for which we assign IP addresses, a very large fraction of posts is generated by just a few IP addresses.

A mere 5% of the 47,630 IP addresses generate over 50% of all posts with assigned IP addresses and 20% of IP addresses generate just over 80% of these posts. While such a high concentration of contributions may appear extreme, it is quite common across many online platforms as documented by Guo et al. (2008, 2009). The degree of concentration is even higher than these numbers might suggest because they only take into account posts with

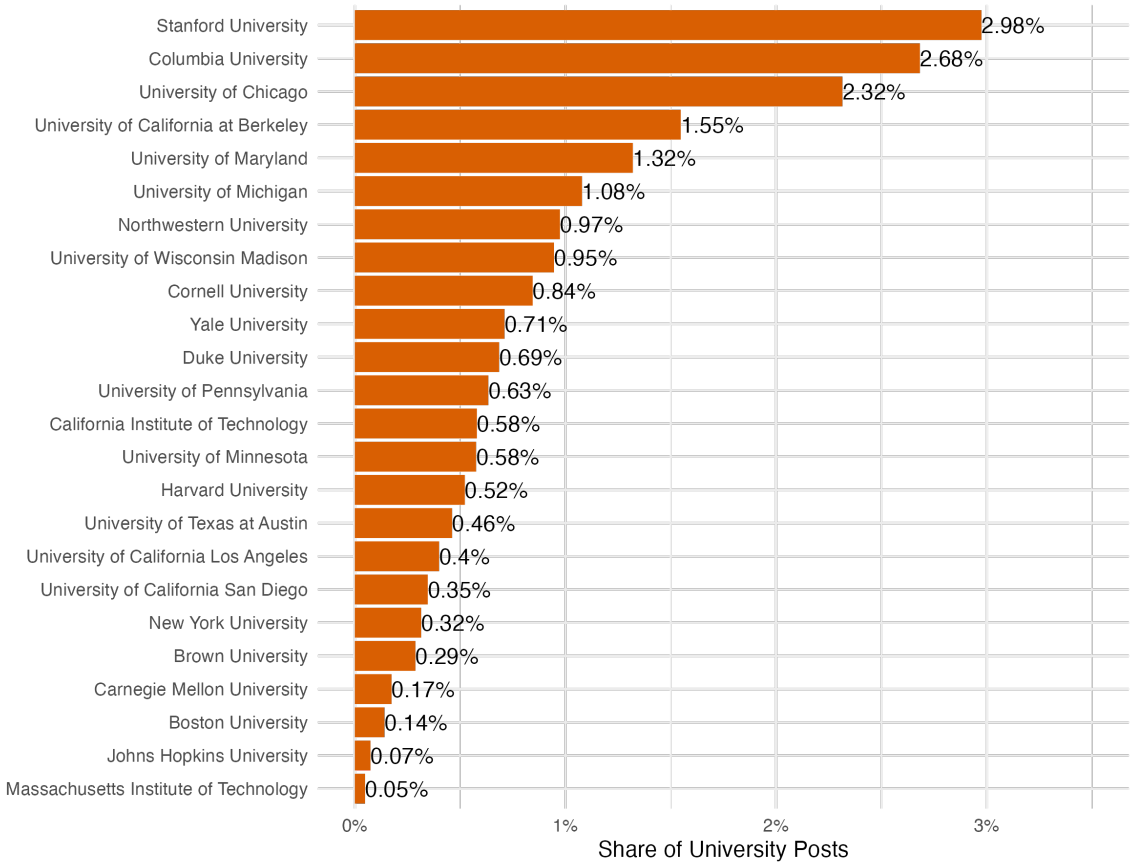


Figure 12: Post share of all university or research institution posts by each of US universities with a top 25 economics department. The figure shows the share of posts accounted for by a given top 25 US university among all posts originating from IP addresses associated with universities or research institutions.

assigned IP addresses. Recall that for 32.1% of assignable posts we do not assign an IP address because the likely IP addresses from which these posts originate do not generate a sufficient number of posts to meet the very conservative identification thresholds we employ. There are thus many more IP addresses with just a few posts each which are not contained in these figures.

Prior research suggests that contributions on online platforms follow neither a power law nor an exponential function, but instead are best approximated by the stretched exponential function (Guo et al., 2008, 2009). This is also the case for EJMR as can be seen in Figure 16 which plots the relationship between IP rank of posters and number of corresponding posts in log-log space. The stretched exponential fits very well up to the point where our assignment procedure stops assigning IP addresses to posts. Loosely speaking, if an IP address posts in fewer than ten topics in the span of a week it will not be assigned to any posts.

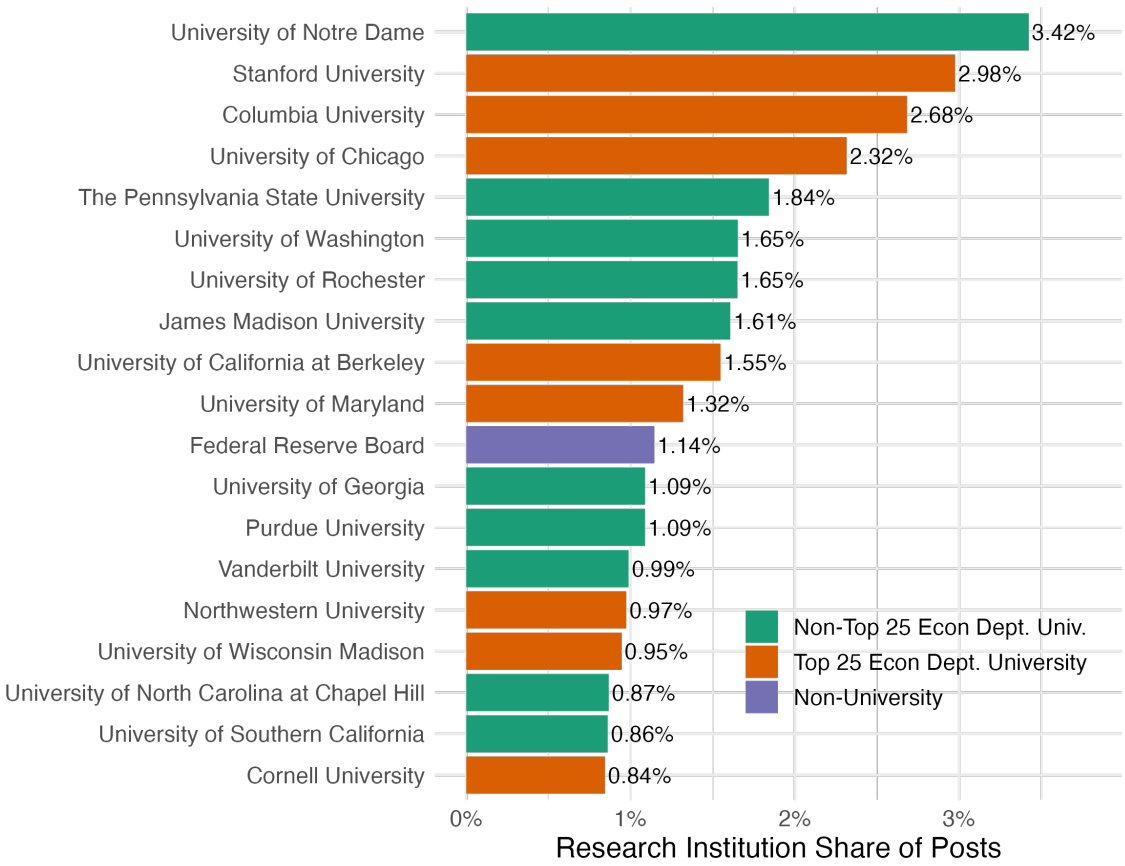


Figure 13: Post share of US university or research institutions. The figure shows the share of posts accounted for by a given US university or research institution among all posts originating from IP addresses associated with US universities or research institutions.

Fitting a stretched exponential distribution to the relationship between IP rank and post count further allows us to estimate how many IP addresses have ever posted on EJMR. We do so by estimating the stretched exponential up to IP rank 40,000 which has 10 posts assigned to it and then projecting out this fitted distribution until the estimated number of posts of an IP address is equal to 1. The estimated total number of posts is 7.4 million and matches the total number of observed posts (7.1 million) quite well. Under this projection there are 582,541 IP addresses which have contributed at least one post to EJMR. Thus, while the vast majority of posts come from just a few thousand IP addresses, our analysis suggests that a very large number of IP addresses has contributed to EJMR over the past decade.

3.5 Content of EJMR Posts

We now turn to analyzing the content of EJMR posts. We focus, in particular, how this content varies across universities and IP addresses.

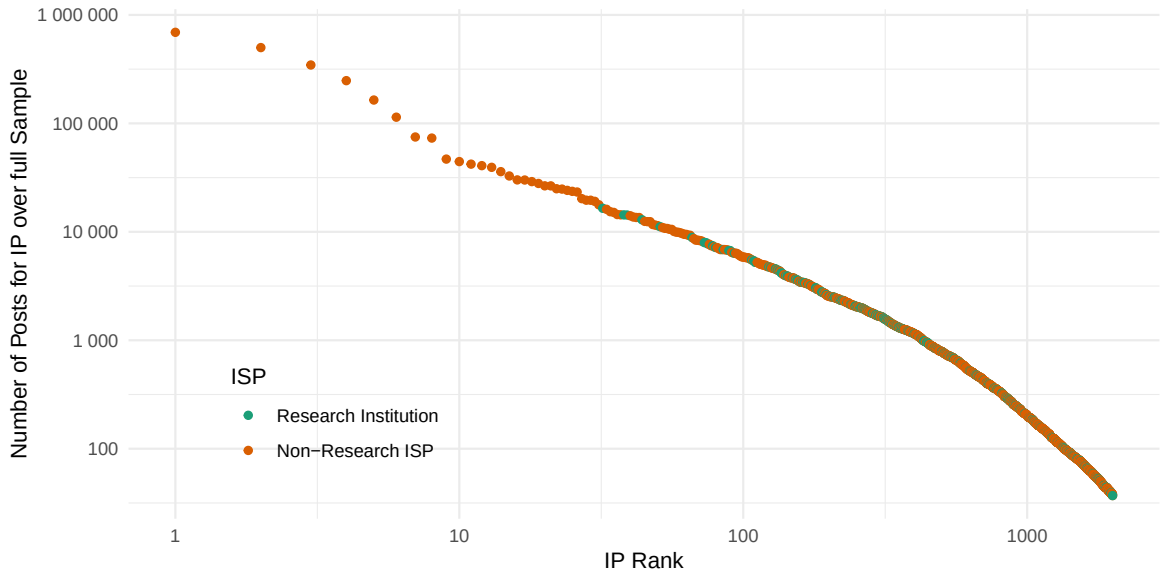


Figure 14: Number of posts by rank of IP address (according to number of posts) split by university and non-university ISP. Green dots indicate that the IP address is a university or research institution.

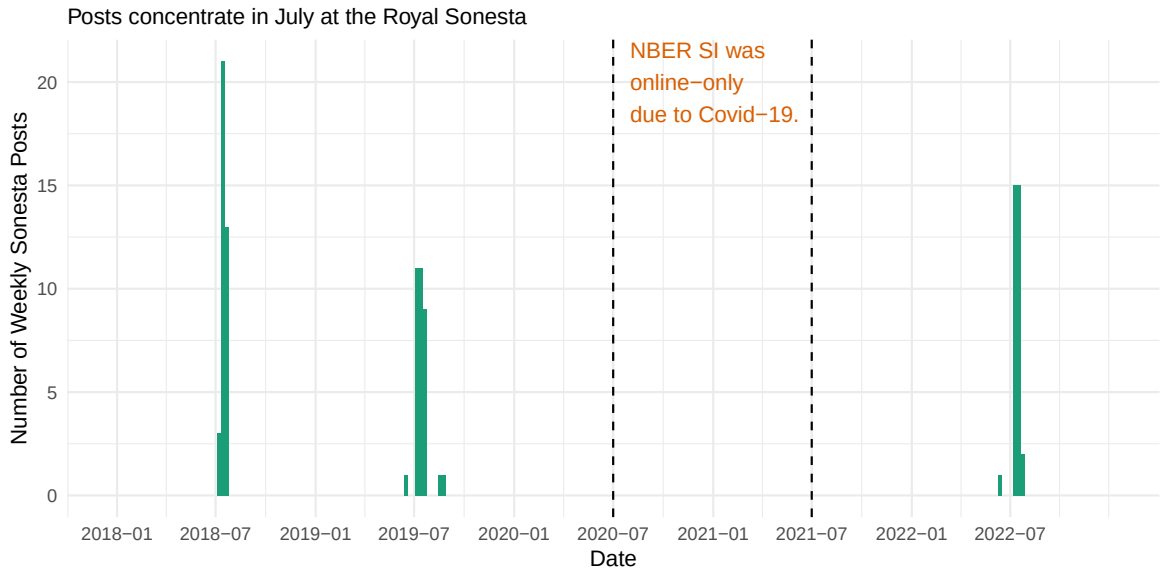


Figure 15: Number of weekly posts from the Royal Sonesta Boston, a hotel in Cambridge, Massachusetts. This hotel is the conference location of the annual NBER Summer Institute, a three-week conference held annually in July where attendance is by invitation only. EJMR posts from this location exclusively occur in July, except in 2020 and 2021 when the NBER Summer Institute was only held virtually rather than in person at the Royal Sonesta Boston.

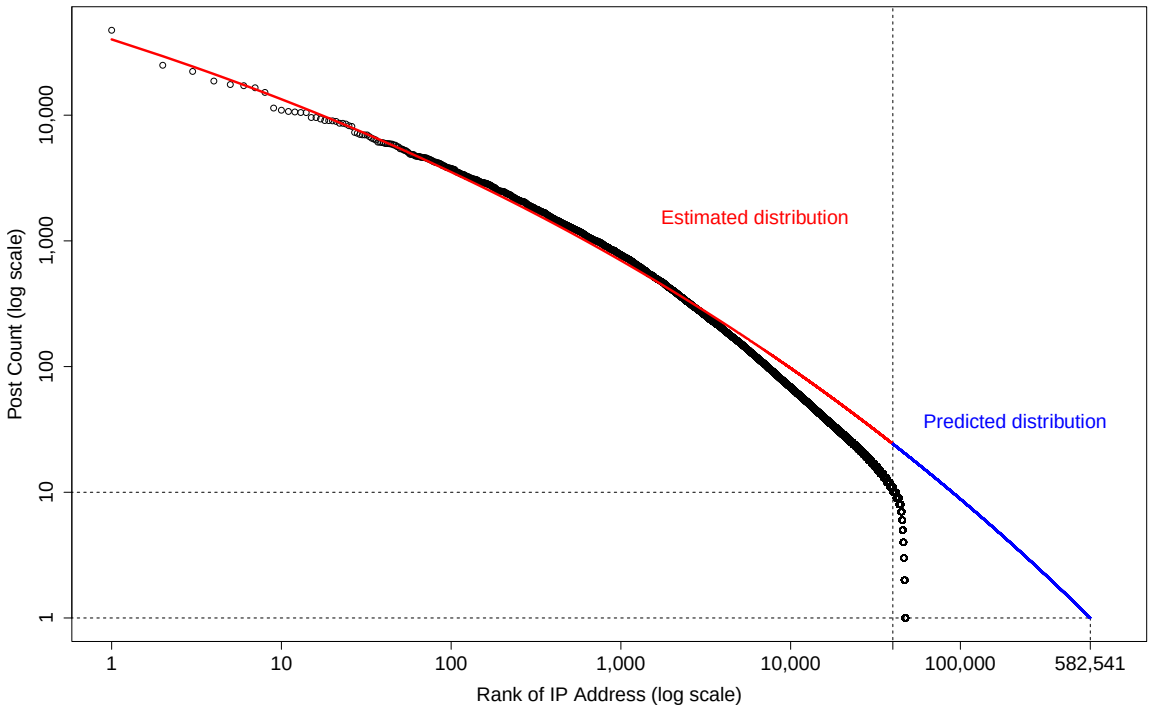


Figure 16: Distribution of posts by IP address rank. The figure shows the number of posts an IP addresses has contributed to EJMR. IP addresses are ordered on the x-axis by the number of posts assigned to them. We estimate a stretched exponential distribution (red line) of posts for the IP addresses ranked 1 to 40,000 and predict the number of posts for all IP addresses with a lower rank (blue). There are 582,541 IP addresses which are predicted to have contributed at least one post to EJMR.

3.5.1 Mentions of Universities

EJMR is intended to be a source of information about the academic job market and of professional news about the economics profession. Posts containing such information may be innocuous and and thus are more likely to come directly from within universities (i.e., from university IP addresses) than toxic posts. A natural way to analyze such information is to investigate how frequently posts from university IP addresses mention their own or other universities, especially for the US universities with the largest number of EJMR posts.

For the ten largest US universities, as measured by the number of EJMR posts originating from IP addresses at theses universities, Table 2 reports the share of posts that mention either the university itself or any other university. Several patterns stand out in this table. First, for all ten universities the largest share of posts mentioning a university is always the share of self-mentions. These self-mention shares are shown on the diagonal and in bold text. This large share of self-mentions may be the result of inside knowledge dissemination.

	Share of posts that mention university									
University ISP	Harvard	MIT	Stanford	Berkeley	UChicago	Yale	NYU	NWU	Columbia	UPenn
Harvard	7.9	9	5.2	1.4	3.7	2	2.1	0.9	1	1.2
MIT	4.7	9.8	6	0.9	2.6	0.4	2.6	1.7	2.1	1.3
Stanford	4.4	6.4	7.4	1.7	4.5	1.2	1.9	1.4	1.3	1.3
UC Berkeley	1.6	3.7	1.7	4.2	1.8	1.2	1.8	0.8	0.6	0.9
UChicago	2.1	4.8	1.4	0.7	8.3	0.8	1.6	0.7	0.5	1
Yale	1.5	3.8	0.9	0.7	1.8	3.4	1.3	0.5	0.4	1
NYU	2.5	4.6	3.1	0.7	1.9	1.1	5.8	1	1.1	2.4
Northwestern	2.5	4.1	1.8	1.1	2.3	1.6	2.7	3.5	0.8	1.1
Columbia	3	4.9	2.3	1.3	2.8	1.6	3.1	1	5	2.4
UPenn	2.2	3.5	1.8	0.9	2.4	1	3.1	1.1	0.3	5.1
Others	1.1	3.8	0.5	0.4	1	0.4	1	0.2	0.3	0.6

Table 2: Share of posts that mention a university from each university ISP. Keyword match for each variable (lower case): Harvard: "harvard|hbs", MIT: "mit|sloan", Stanford: "stanford", Berkeley: "berkeley|haas", UChicago: "uchicago|university of chicago|chicago|booth", Yale: "yale", NYU: "nyu|stern", Northwestern: "northwestern|kellogg", Columbia: "columbia", UPenn: "upenn|penn|wharton"

Second, mentions of other universities tend to decline by university rank. For example, the share of posts mentioning Harvard is larger than the share of posts mentioning Columbia, Northwestern, and UPenn for every single university ISP except, of course, the own university ISP. Third, even among these institutions MIT stands out. Posts from other university ISPs shown in the last row of Table 2 mention MIT almost four times more often than any of the other top 10 institutions.

3.5.2 Misogyny, Toxicity, and Hate Speech

As described in Section 2.5, we deobfuscated the content and then used a number of transformer-based machine learning models to classify EJMR posts by sentiment, misogyny, and toxicity. Each of these transformer models is best-in-class yet nonetheless imperfect. We also re-created all the word-count measures used in the seminal study of Wu (2020).

EJMR employs automatic moderation that deletes profane speech but does not prohibit toxic speech more generally. For example, users are not allowed to write "fuck" but they can (and do) write "these brahmin jeets are replacing the juifs in terms of financial engineering excellence. when do we start building the crematoriums?"[17] Clearly, such posts in our data were also not removed by EJMR's human moderators.

Overall, the share of EJMR posts classified as misogynistic, toxic, or hate speech is 2.5%, 10%, and 2.5% respectively. These shares have remained relatively constant since the inception of EJMR with some mild seasonality over the course of each year with a higher share of misogynistic, toxic, or hate speech posts during the summer months. Across IP addresses the concentration of posts containing such offensive content is generally higher than that of other non-problematic posts. 20% of identified IP addresses generate 85% of all toxic posts with assigned IP addresses. Moreover, posts in the Off-Topic/Non-Econ forums are substantially more likely to be misogynistic (almost 4%) or toxic (roughly 12%) than those in the Economics (2% and 10%) or Job Market Rumors (2% and 8%) forums. However, even in the Job Market Rumors forums in which discussion focuses on the academic job market, approximately 2% and 8% of all posts are classified as misogynistic or toxic.

In addition, there is also considerable heterogeneity of the share of problematic posts across universities. Figure 18 reports a scatterplot of the total number of EJMR posts and combined share of toxic, misogynistic, or hate speech posts by university for all university ISPs. As is evident from the figure, usage (both problematic and not) of EJMR, is widespread throughout the economics profession and not limited to any particular subset of institutions.

Although our results so far show that many posts on EJMR contain problematic content,

[17] "Jews" is variously obfuscated on EJMR, usually as merely "Js" while "pajeet" and "jeet" are racial epithets used in reference to persons of South Asian decent.

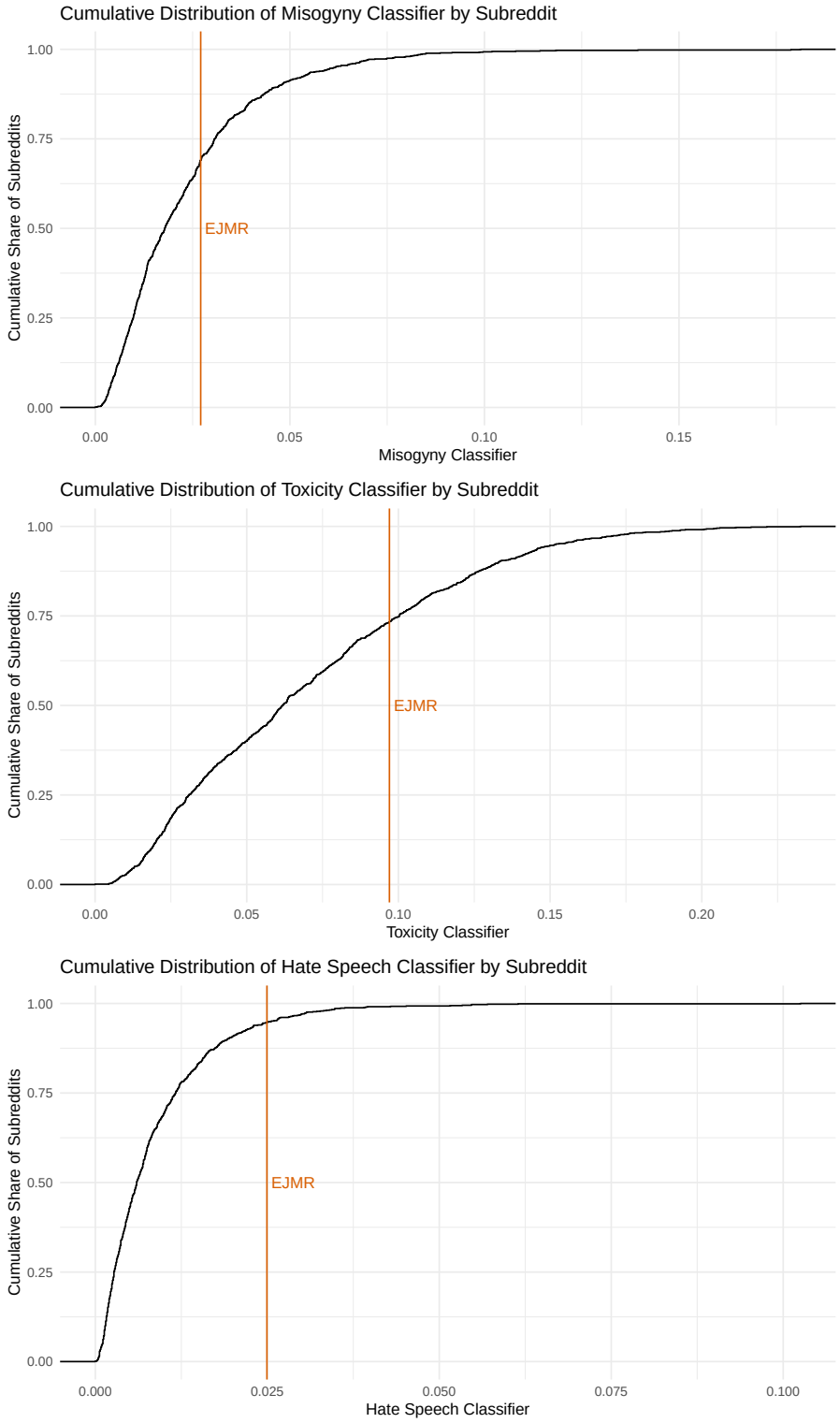


Figure 17: Cumulative distributions of the share of posts containing misogyny, toxicity, and hate speech for the 1,000 most popular subreddits. The respective positions of EJMR in these distributions are indicated by the orange lines.

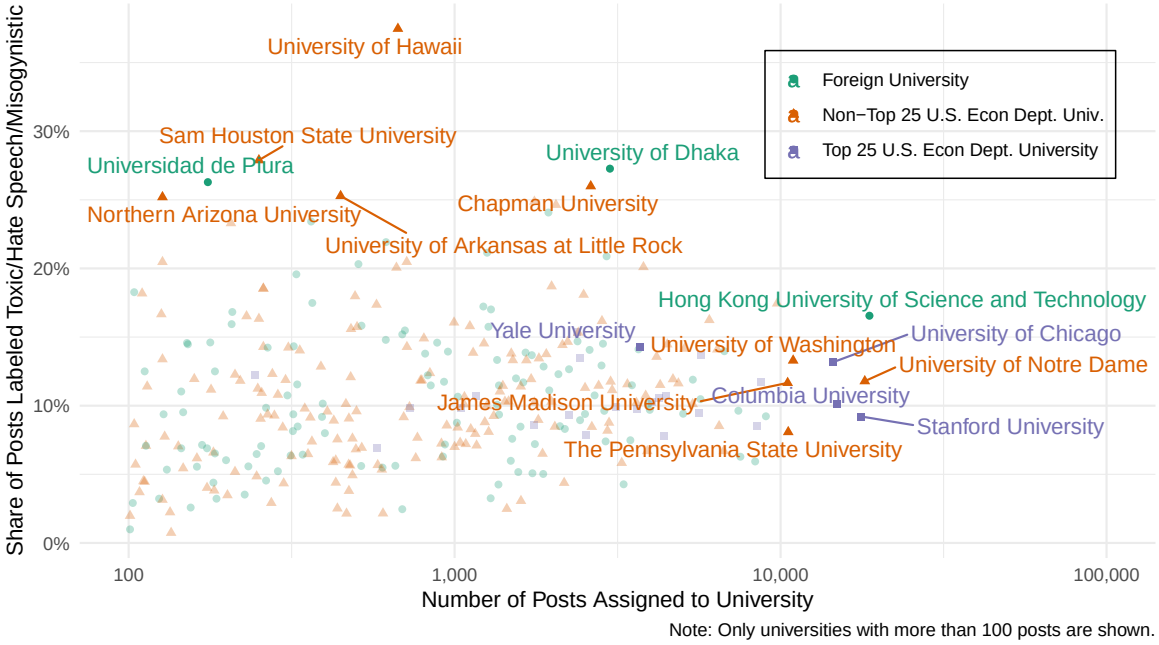


Figure 18: Total number of EJMR posts assigned to university and share of posts labeled toxic, misogynistic, or hate speech by university. The figure shows a scatterplot of the total number of EJMR posts and share of toxic, misogynistic, or hate speech posts by university for all university ISPs. US universities with an economics department ranked in the top 25 are marked in purple, US universities with an economics department ranked outside the top 25 are marked in orange, and non-US universities are marked in green.

it is less clear whether the proportion of this content on EJMR is particularly egregious when compared to less professional settings. To evaluate whether the aforementioned shares of problematic content are relatively high or low, it is instructive to compare them to content on other internet platforms in which users can contribute posts while remaining relatively anonymous. As a validation, we therefore compare EJMR posts to Reddit posts.

Figure 17 reports the cumulative distribution of the share of posts containing misogyny, toxicity, and hate speech classifiers for the 1,000 most popular subreddits. The respective positions of EJMR in these distributions are indicated by the orange lines. Compared to these subreddits, EJMR ranks at the 69th percentile of misogyny, the 73rd percentile of toxicity, and the 95th percentile of hate speech. Taken together, this suggests that even compared to other anonymous internet speech in a strictly non-professional setting such as Reddit, EJMR contains markedly more toxic material. This conclusion becomes even clearer when comparing EJMR to specific subreddits dedicated to economics or finance. Among the 1,000 most popular subreddits, the three most similar in terms of general topic are r/badeconomics, r/finance, and r/AskEconomics. Figure 19 shows that EJMR has more posts labeled misogynistic, toxic, and hate speech than these related subreddits.

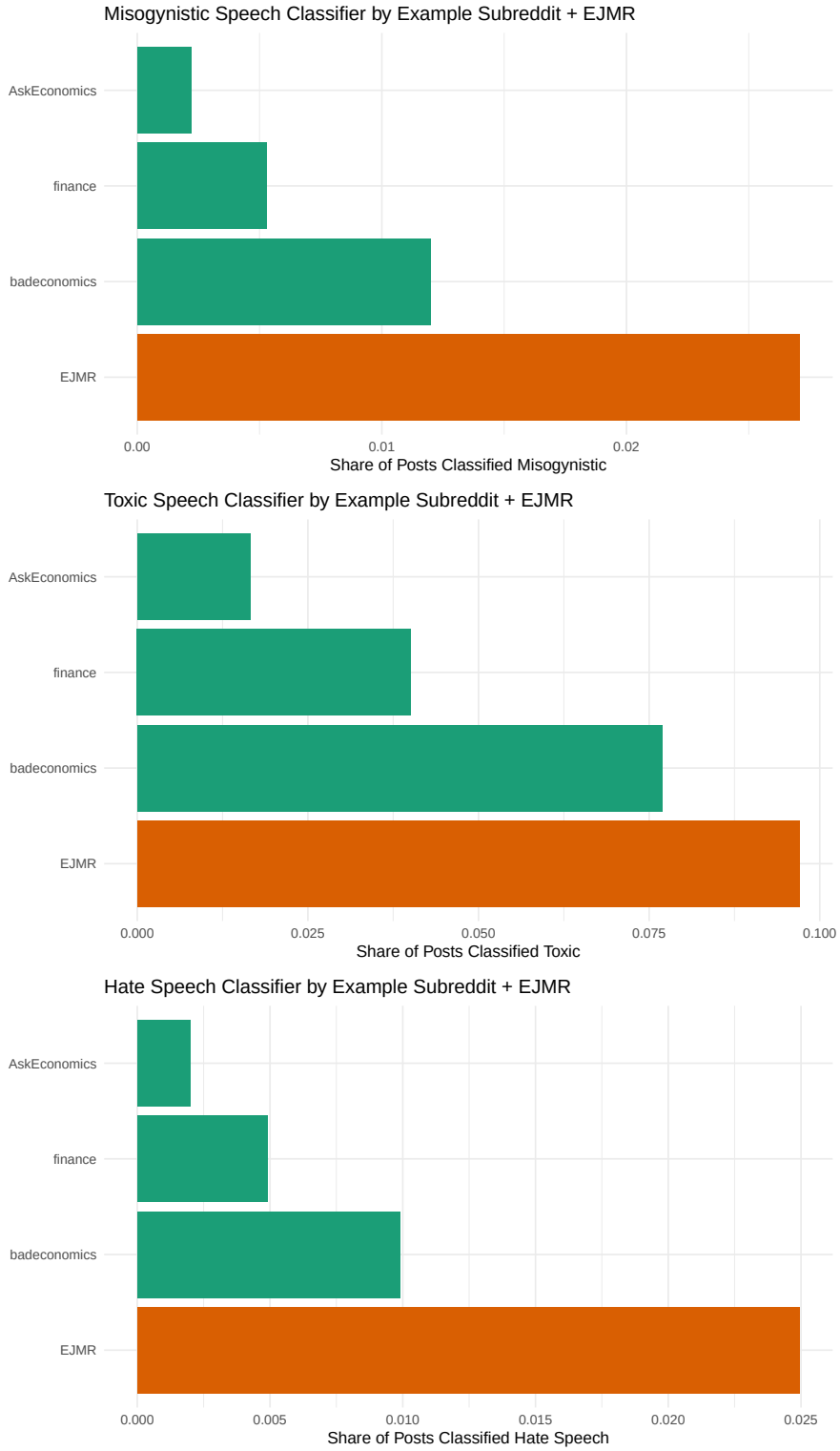


Figure 19: Share of posts classified as containing misogyny, toxicity, and hate speech for r/AskEconomics, r/finance, r/badeconomics, and EJMR.

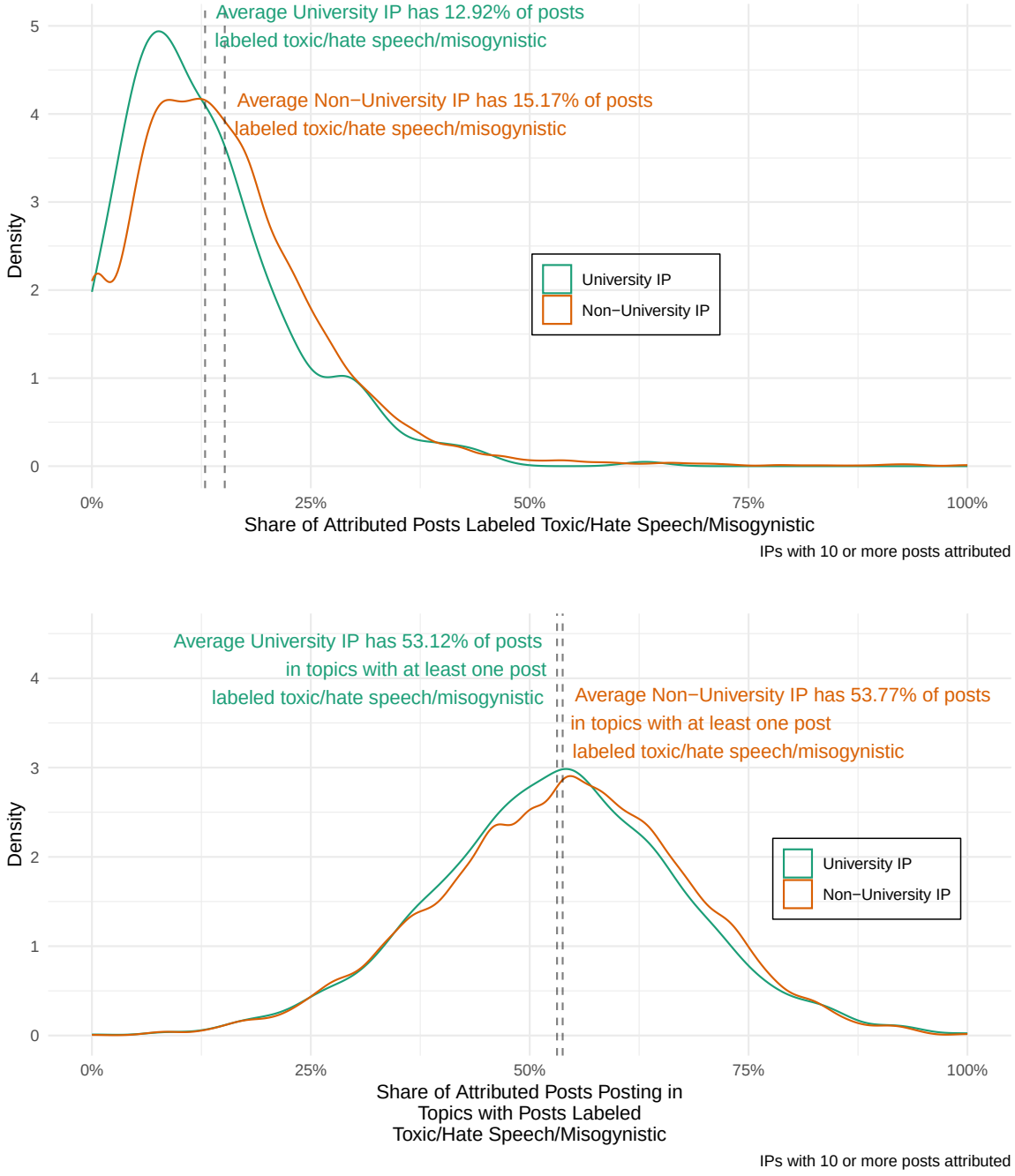


Figure 20: Distribution of misogynistic, toxic, and hate speech across IPs for university and non-university IP addresses. The figure plots density functions of the share of misogynistic, toxic, and hate speech posts across IP addresses (top panel) and density functions of the share of posts in topics with at least one post classified as misogynistic, toxic, and hate speech (bottom panel). The green lines shows the density functions for university IP addresses and the orange lines for non-university IP addresses. Only IP addresses with 10 or more posts attributed to them are shown in the figure.

Perhaps the most concerning aspect of EJMR is that, in contrast to Reddit, it features specific commentary (and attacks) on economists who are, with a few notable exceptions, not public figures such as singers, dancers, actors, and politicians who seek out public attention. This is important because private figures are more vulnerable and much less likely to have the resources to litigate defamatory content about them. They are also less likely to receive high damages awards for reputational injury assuming they could find an attorney to take the case on a contingency fee. Public figures, in contrast, have the resources to hire lawyers and generally can use their access the media to rebut defamatory statements without assistance from the courts.

One might expect that EJMR users exhibit greater inhibition to contribute offensive content when connected through their work compared to when connected at home. If this were true, we should observe that posts from university IP addresses are less likely to be toxic on average than posts from non-university IP addresses. The top panel of Figure 20 shows that this is indeed the case. The average university IP has 12.9% of posts labeled as misogynistic, toxic, or hate speech whereas that share for the average non-university IP is equal to 15.2%. This pattern also holds across all three groups of EJMR subforums. For each group of subforums, posts coming from university IP addresses are roughly three percentage points less likely to be problematic than posts originating from non-university IP addresses. This pattern is slightly less pronounced for the share of posts that IP addresses contribute in topics with at least one post classified as misogynistic, toxic, and hate speech as can be seen in the bottom panel of Figure 20. The orange university density is only slightly to the right of the green non-university density. EJMR contributors from university and non-university IP addresses thus appear to be equally willing to engage with problematic content. And they do so at very high rates given that the average IP address has over 50% of its posts in topics with at least one post that is misogynistic, toxic, or hate speech.

Finally, among the top 10 IP addresses with the highest number of toxic posts, there is not a single one from a university IP address. However, among the top 10 university IP addresses with the highest number of toxic posts, there are several from leading universities and economics departments including the University of Chicago, the University of Rochester, the University of Washington, and University College London as well as from less prominent institutions such as Virginia Wesleyan University and Lingnan University. This fact pattern again underscores the diversity and pervasiveness of toxic speech on EJMR and in the economics profession.

Our analysis so far relies on simple comparisons of average shares of problematic posts on EJMR and Reddit. However, in addition to the vastly different target group comparing these two platforms might be complicated by the fact that the posts are different in their length

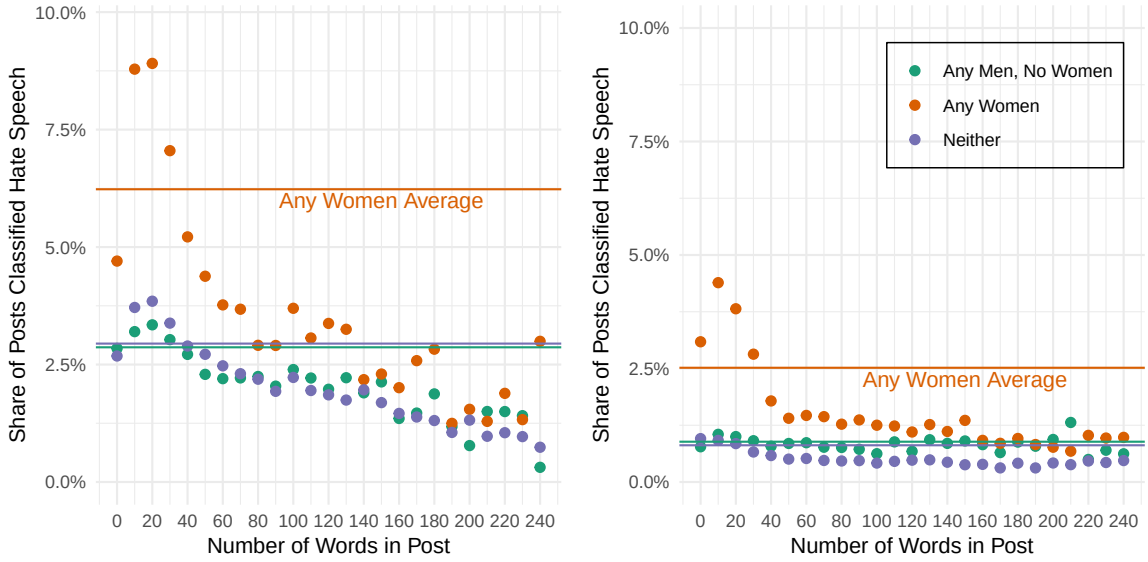


Figure 21: Share of EJMR and Reddit posts classified as hate speech by number of words. The figure shows the share of posts classified as hate speech by number of words on EJMR (left panel) and Reddit (right panel) for women (orange), men (green), and neither (purple).

and form. For example, it is possible that our toxicity classifiers may capture features that differ across length of post. We therefore examine how our classifiers vary by length of post (i.e., word count). We find that on both platforms short messages are more likely to be toxic than long messages. More importantly, for any length of a post, EJMR posts are always more likely to be labeled as misogynistic, toxic, and hate speech than Reddit on average.

One concern about the classifiers we use is that they may be labeling posts with noise. As a result, it is natural to ask how much of the difference between the platforms is due to the difference in toxicity, the difference in noise, or both? We use the response in toxicity to the mentions of women to identify the difference in toxicity across platforms more sharply. The idea is that comparing the toxicity of posts that mention women to those that mention neither men nor women will capture the change in true toxicity rather than changes in errors. To do so, we classify posts containing female words ("she", "her", "hers", "woman", "girl", "gal") as mentioning women and posts containing male words ("he", "him", "his", "man", "boy", "guy") as mentioning men. Figure 21 shows the share of posts classified as hate speech by the number of words contained in the post on EJMR (left panel) and Reddit (right panel) for women (orange), men (green), and neither men nor women (purple). On both platforms, posts mentioning women are substantially more likely to contain hate speech and the same is true for misogyny and toxicity.

To be even more conservative, we compare the difference between posts mentioning women and those mentioning neither men nor women to the difference between posts mentioning

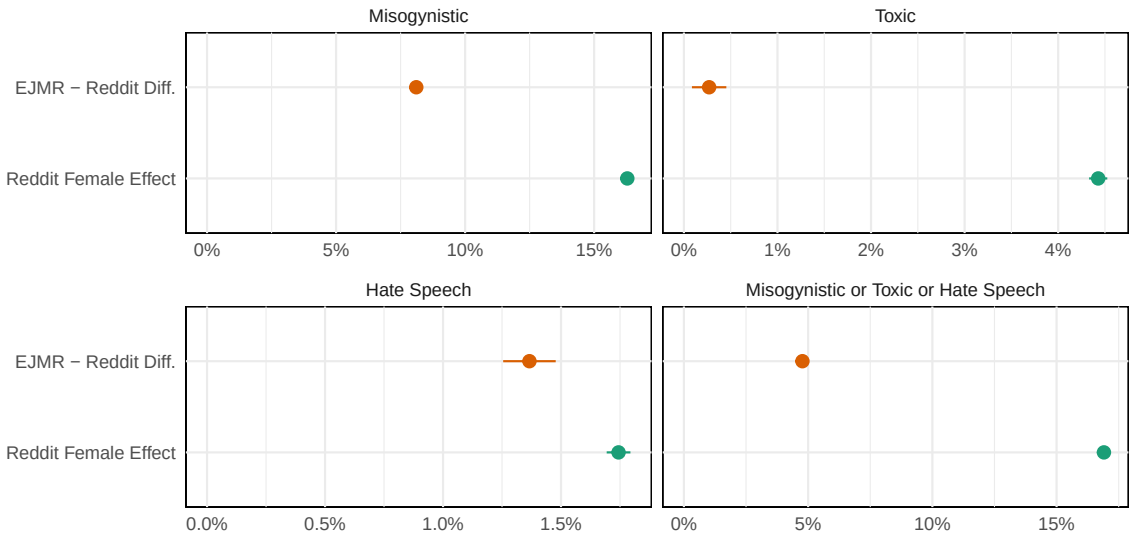


Figure 22: Point estimates and standard error bars for differences in misogyny, toxicity, hate speech, and combined offensive speech. The Reddit female effect estimates (green) show that compared to Reddit posts mentioning men, Reddit posts mentioning women are substantially more likely to be misogynistic, toxic, and hate speech. This pattern is even more pronounced on EJMR where the *additional* effect relative to Reddit (orange) is sizable for misogyny and hate speech, but quite small for toxicity.

men and those mentioning neither men nor women. We consider this gap to measure the "true" toxicity of the platform because it quantifies how much more toxic a platform is when women are mentioned compared to men. We then estimate and compare this difference in differences for both EJMR and Reddit and report the point estimates and standard error bars in Figure 22. The Reddit female effect estimates (green) show that compared to Reddit posts mentioning men, Reddit posts mentioning women are substantially more likely to be misogynistic, toxic, and hate speech. However, this effect is even more pronounced on EJMR where posts mentioning women are *even more likely* to include problematic content when compared to Reddit (orange). Across the board, we find that mentions of women attract even more misogyny (top-left panel) and hate speech (bottom-left panel) on EJMR than on Reddit, but a relatively similar level of toxicity (top-right panel).

3.5.3 Network Relationships

Our IP identification allows us to analyze the content of EJMR posts beyond the relatively simple aggregation of locations and universities. Figure 23 in the appendix shows the linguistic patterns of IP addresses contributing to EJMR. The figure shows one point for each IP address that contributed to EJMR in our data (minus a few not meeting the criteria below). Proximity

in the figure indicates linguistic similarity. As is evident from the figure, even among English speakers there are clusters of linguistically similar users.

In addition, Figure 24 in the appendix gives a glimpse of the interactions between different IP addresses. Each vertex is an IP address in this graph. Edges between vertices indicate that two IP addresses often posted in the same topics and the color indicates the mean year in which an IP address was active. The graph is laid out using the ForceAtlas2 algorithm (Jacomy et al., 2014). The figure shows that IP-to-IP interactions do not occur over long stretches of time. Contributors change IPs and the popular topics on EJMR change, leading to the march of colors across time.

4 Conclusion

In this paper we analyzed the behavior of posters on EJMR, a popular online platform for economists that allows users to read and post anonymously. Using only publicly available data we showed that the statistical properties of the scheme by which EJMR assigned usernames to posts until May 2023, identify the IP addresses from which most posts were made. To recover these IP addresses we employed a multi-step procedure. First, we developed GPU-based software to quickly compute the SHA-1 hashes used for the username allocation algorithm on EJMR. Second, we measured which IP addresses occur particularly often in a narrow time window and used the uniformity property of the SHA-1 hash to test whether these IP addresses appear more often than would likely occur by chance.

We recovered 47,630 distinct IP addresses of EJMR posters and attributed them to 66.1% of the roughly 7 million posts made over the past 12 years. Based on the geographic location of these IP addresses, we showed that the majority of posts come from large cities (and also smaller cities with elite universities) in the US and other developed countries with leading research institutions such as Canada, the United Kingdom, Germany, France, and Hong Kong. We further showed that posting on EJMR is pervasive throughout the economics profession including all top-ranked universities in the United States. A substantial number of posts also come from government agencies, companies, and non-profit organizations employing economists as well as universities around the world. Finally, we showed that EJMR contains much problematic content that violates the professional conduct code for economists, particularly in posts that target women.

Taken together, our paper provides further evidence of a toxic environment that is pervasive at all echelons of the economics profession, including, but not limited to, its most elite institutions.

References

Adamic, Lada A and Eytan Adar, "Friends and neighbors on the Web," *Social Networks*, 2003, *25* (3), 211–230.

Antecol, Heather, Kelly Bedard, and Jenna Stearns, "Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies?," *American Economic Review*, 2018, *108* (9), 2420–2441.

Attanasio, Giuseppe, Debora Nozza, Dirk Hovy, and Elena Baralis, "Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists," in "Findings of the Association for Computational Linguistics: ACL 2022" Association for Computational Linguistics Dublin, Ireland May 2022, pp. 1105–1119.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn, "The Pushshift Reddit Dataset," 2020.

Bayer, Amanda and Cecilia Elena Rouse, "Diversity in the economics profession: A new attack on an old problem," *Journal of Economic Perspectives*, 2016, *30* (4), 221–242.

Benjamini, Yoav and Yosef Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995, *57* (1), 289–300.

Biscarri, William, Sihai Dave Zhao, and Robert J. Brunner, "A simple and fast method for computing the Poisson binomial distribution function," *Computational Statistics & Data Analysis*, 2018, *122*, 92–100.

Blanchard, Olivier, "The Economics Job Market Rumors Site Needs to Clean Up Its Act," *Peterson Institute for International Economics*, 2017.

Choquette, Jack, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky, "Nvidia a100 tensor core gpu: Performance and innovation," *IEEE Micro*, 2021, *41* (2), 29–35.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, 2021, *118* (9), e2023301118.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman, "Power-law distributions in empirical data," *SIAM review*, 2009, *51* (4), 661–703.

Cormen, Thomas H, Charles E Leiserson, Ronald L Rivest, and Clifford Stein, *Introduction to Algorithms*, MIT Press, 2022.

Cotton, Michelle, Leo Vegoda, Ron Bonica, and Tim Chown, "Special Use IPv4 Addresses," Internet Engineering Task Force (IETF) 2010. Available from: https://www.rfc-editor.org/rfc/rfc5735.txt.

Damgård, Ivan, "A design principle for hash functions," in "Crypto," Vol. 89 1990, pp. 416–427.

Dupas, Pascaline, Alicia Sasser Modestino, Muriel Niederle, Justin Wolfers, and The Seminar Dynamics Collective, "Gender and the dynamics of economics seminars," *NBER Working Paper*, 2021.

Ferguson, Niels, Bruce Schneier, and Tadayoshi Kohno, *Cryptography Engineering* 03 2010.

Guo, Lei, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong (Eric) Zhao, "Analyzing Patterns of User Content Generation in Online Social Networks," in "Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" KDD '09 Association for Computing Machinery New York, NY, USA 2009, p. 369–378.

Guo, Lei, Enhua Tan, Songqing Chen, Zhen Xiao, and Xiaodong Zhang, "The stretched exponential distribution of internet media access patterns," in "Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing" 2008, pp. 283–294.

Hartvigsen, Thomas, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar, "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection," in "Proceedings of the 60th Annual Meeting of the Association of Computational Linguistics" 2022.

Hengel, Erin, "Publishing while female: Are women held to higher standards? Evidence from peer review," *Economic Journal*, 2022, *132* (648), 2951–2991.

Hochberg, Yosef and Ajit C Tamhane, *Multiple comparison procedures*, John Wiley & Sons, Inc., 1987.

Hugging Face, "distilbert-base-uncased-finetuned-sst-2-english," https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english 2023. [Accessed 21-Jun-2023].

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software," *PLoS One*, June 2014, *9* (6), e98679.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *CoRR*, 2019, *abs/1907.11692*.

Livadariu, Ioana, Karyn Benson, Ahmed Elmokashfi, Amogh Dhamdhere, and Alberto Dainotti, "Inferring Carrier-Grade NAT Deployment in the Wild," in "IEEE INFOCOM 2018 - IEEE Conference on Computer Communications" 2018, pp. 2249–2257.

Lowrey, Annie, "Harassment in Economics Doesn't Stay in Economics," *The Atlantic*, 2022.

Lundberg, Shelly and Jenna Stearns, "Women in economics: Stalled progress," *Journal of Economic Perspectives*, 2019, *33* (1), 3–22.

McInnes, Leland, John Healy, and James Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2020.

McInnes, Leland, John Healy, and Steve Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, 2017, *2* (11), 205.

Mishra, Vikas, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Martin Lopatka, "Don't Count Me Out: On the Relevance of IP Address In The Tracking Ecosystem," in "Proceedings of The Web Conference 2020" WWW '20 Association for Computing Machinery New York, NY, USA 2020, p. 808–815.

Mochimo Cryptocurrency Engine, "Mochimo GitHub Repository," 2023.

Motara, Yusuf Moosa and Barry Irwin, "Sha-1 and the strict avalanche criterion," in "2016 Information security for South Africa (ISSA)" IEEE 2016, pp. 35–40.

Nakamoto, Satoshi, "Bitcoin: A Peer-to-Peer Electronic Cash System," Dec 2008. Accessed: 2015-07-01.

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *CoRR*, 2019, *abs/1908.10084*.

Romer, David, "Evidence of a toxic environment for women in economics," in Justin Wolfers, ed., *Evidence of a toxic environment for women in economics*, New York Times, 2017.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *ArXiv*, 2019, *abs/1910.01108*.

Saxon, James and Nick Feamster, "GPS-Based Geolocation of Consumer IP Addresses," in Oliver Hohlfeld, Giovane Moura, and Cristel Pelsser, eds., *Passive and Active Measurement*, Springer International Publishing Cham 2022, pp. 122–151.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in "Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing" Association for Computational Linguistics Seattle, Washington, USA October 2013, pp. 1631–1642.

Spiegel, H. W., "Jacob Viner (1892–1970)," in J. Eatwell, M. Milgate, and P. Newman, eds., *The New Palgrave: A Dictionary of Economics*, Vol. IV, London: Macmillan, 1987, p. 812–14.

Spinellis, Diomidis, "Git," *IEEE software*, 2012, *29* (3), 100–101.

Stahl, Peter M., "pemistahl/lingua-py: The most accurate natural language detection library for Python, suitable for long and short text alike — github.com," https://github.com/pemistahl/lingua-py 2023. [Accessed 21-Jun-2023].

Standard, Secure Hash, "FIPS Pub 180-1," *National Institute of Standards and Technology*, 1995, *17* (180), 15.

Tang, Wenpin and Fengmin Tang, "The Poisson Binomial Distribution—Old & New," *Statistical Science*, 2023, *38* (1), 108–119.

Taylor, Kate, "'4chan for economists' is melting down as racist, sexist anonymous posts are linked to Harvard, Yale, and other top institutions," *Business Insider*, 2023.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023.

Wolfers, Justin, "Evidence of a toxic environment for women in economics," *New York Times*, 2017.

Wordpress Foundation, "bbPress," 2023.

Wu, Alice H, "Gendered Language on the Economics Job Market Rumors Forum," *AEA Papers and Proceedings*, 2018, *108*, 175–79.

Wu, Alice H, "Gender Bias Among Professionals: An Identity-based Interpretation," *Review of Economics and Statistics*, 2020, *102* (5), 867–880.

A Frequently Asked Questions

There is a some misinformation about this manuscript online, particularly on EJMR. It is possible that you have encountered that misinformation and it seems prudent for us to address it in this FAQ.

A.1 I read on EJMR that this is a "hack." Is that true?

No. Our study uses *only* publicly available pages on EJMR, the same pages viewed by other EJMR users and indexed by search engines such as Google, Yandex, Baidu, Bing, and Archive.org. At no point did we access any non-public pages, hidden URLs, or APIs. Every page of EJMR we used in our study was in a chain of links from the EJMR homepage. Every page permitted access from EJMR's robots.txt, EULA (none), and terms of service (none). Furthermore, every single page from EJMR used in this study contained advertisements. All that is to say, everything in this study was visible to us as ordinary consumers of the ordinary EJMR content designed for public consumption. What we present in this paper is merely a statistical analysis of that ordinary, publicly available content, particularly the usernames shown on EJMR until May 2023.

A.2 Did you abuse EJMR's servers in any manner?

No. Our study used about 350,000 unique HTML pages from EJMR. According to Semrush Traffic Analytics, at the time of our scrape, EJMR received 824,000 monthly visits with an average of 12.5 pages per visit which amounted to 10.3 million monthly pageviews. Not only is our total number of pages a small fraction of EJMR's traffic, but we also used a commercial, third-party web indexer to obtain our data. The indexer we used is a major, well-known company that clearly labels its user agent, respects robots.txt directives, and meters its requests to avoid imposition on web administrators.

A.3 Are you "doxing" people?

No. Our study fits squarely in the rich tradition of scholarship on hate speech, harassment, and online communities. Our study does not dox any users. The paper reveals a single IP address, which is provided specifically in response to a public million-dollar prize offered by EJMR's owner for identifying the IP address of one particular post. We hope to collect and donate that prize to a suitable charity.

B Additional Figures

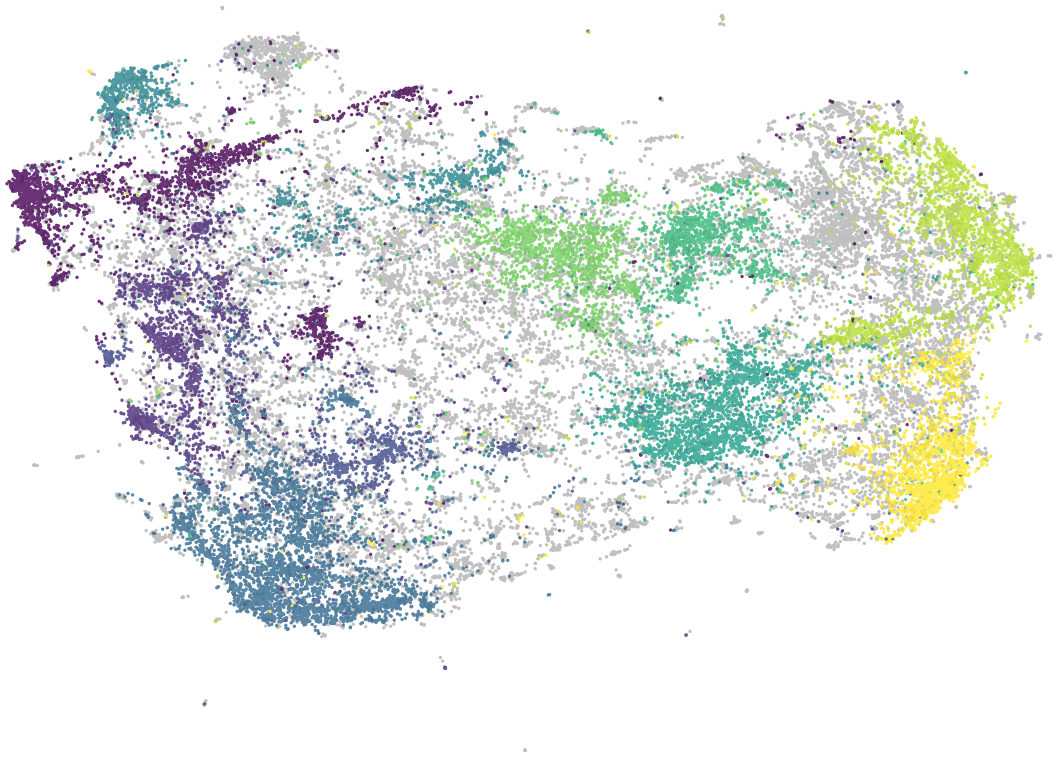


Figure 23: Linguistic patterns of EJMR contributing IPs. The chart shows one point for each IP address that contributed to EJMR in our data set (minus a few not meeting the criteria below). Proximity in the figure indicates linguistic similarity. To make this graph each post that was in English, at least 25 words, and at least 100 characters was embedded into a vector space using SBERT (Reimers and Gurevych, 2019). These vectors were averaged for each IP address and projected onto a two-dimensional manifold using umap (McInnes et al., 2020). Clusters were created using HDBSCAN (McInnes et al., 2017). Unclustered IP addresses are in grey. The figure shows that substantial clusters of linguistically similar users exist.

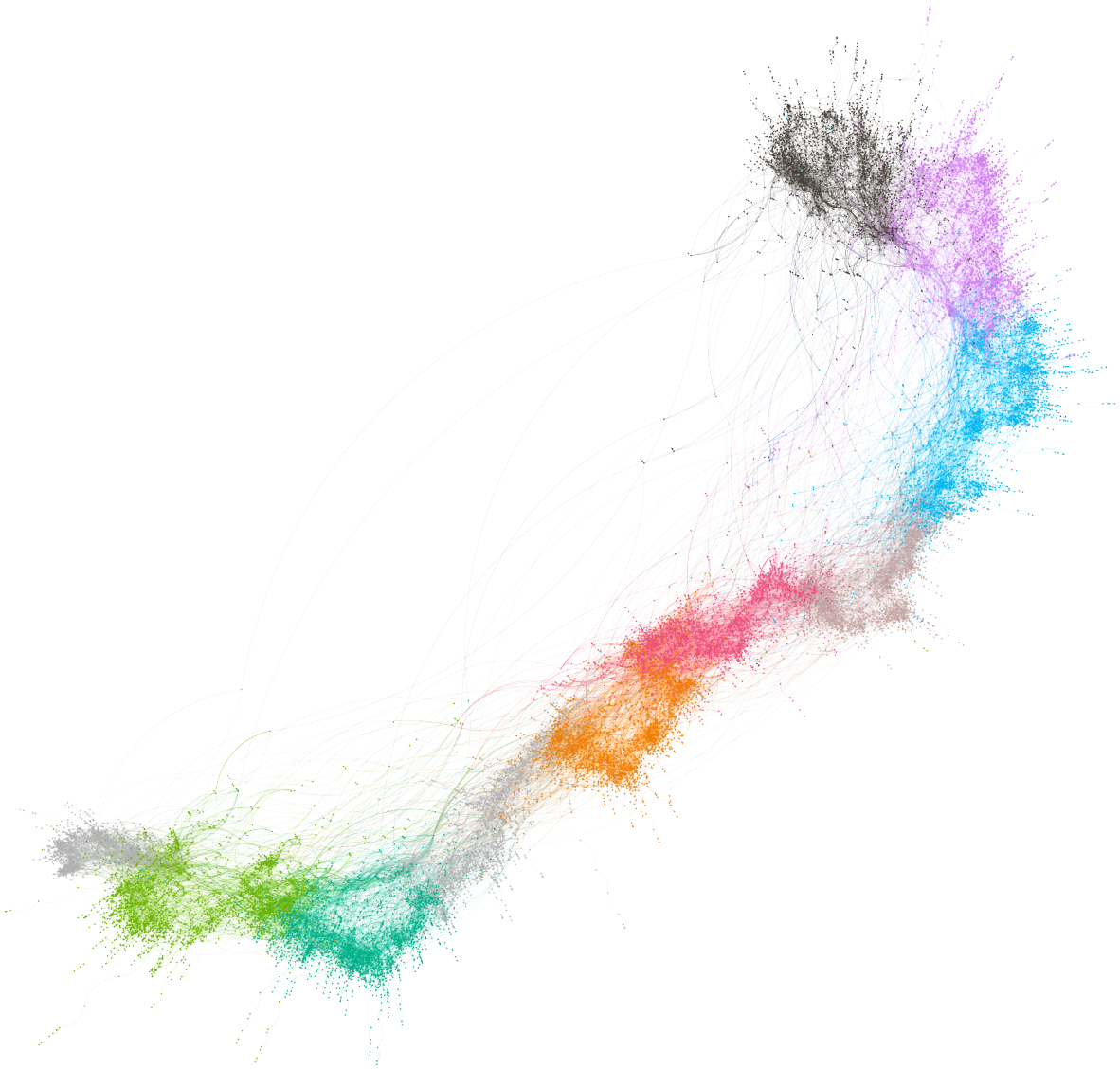


Figure 24: Interactions between IP addresses. The graph includes most of the IP addresses identified in our study with one vertex per IP address. Edges between vertices indicate that two IP addresses often posted in the same topics on EJMR. Color indicates mean year in which an IP address was active. To make this graph we computed the pairwise Adamic-Adar similarity metric between users (Adamic and Adar, 2003) based on shared topics. This metric is conceptually similar to a TF-IDF metric in that IPs receive a high pairwise score if they posted together in topics that are not generally popular. We dropped the top 10 percent of highly active IP addresses and all edges below the 90th percentile of weight. The graph is laid out using the ForceAtlas2 algorithm (Jacomy et al., 2014). The figure shows that IP-to-IP interactions do not occur over long stretches of time. That is, contributors change IPs and the popular topics on EJMR change, leading to the march of colors across time.

C Posts by University ISP

Table 3: Total number of EJMR posts, share of posts labeled toxic, misogynistic, or hate speech, and share of posts labeled as containing positive or negative sentiment by university for all universities with more than 100 posts. Some universities are shown with multiple entries as their discipline-specific subunits if the ISP provides such information. These subunits are appropriately aggregated at the university level in Figure 18.

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
HKUST	HK	0.16	0.14	0.04	0.03	0.53	0.22	18934
University of Notre Dame	US	0.12	0.09	0.02	0.03	0.64	0.25	18498
Stanford University	US	0.09	0.08	0.01	0.02	0.53	0.27	17813
Columbia University	US	0.10	0.08	0.02	0.02	0.57	0.29	14970
University of Chicago	US	0.13	0.11	0.02	0.03	0.52	0.24	14597
University of Washington	US	0.13	0.12	0.02	0.02	0.50	0.21	11089
Pennsylvania State University	US	0.08	0.06	0.01	0.02	0.59	0.24	10723
James Madison University	US	0.12	0.10	0.02	0.01	0.54	0.19	10598
University of Rochester	US	0.17	0.16	0.04	0.02	0.57	0.23	9802
University of Oxford	GB	0.09	0.08	0.02	0.02	0.47	0.24	9296
University of California at Berkeley	US	0.12	0.09	0.04	0.01	0.59	0.24	8784
University of Maryland	US	0.08	0.07	0.01	0.02	0.50	0.23	8638
ROISNII	JP	0.06	0.04	0.01	0.01	0.54	0.31	8408
Fraser Institute	CA	0.06	0.05	0.01	0.02	0.50	0.20	7745
Cardiff University	GB	0.10	0.07	0.02	0.02	0.62	0.31	7435
Rouen Management School	FR	0.14	0.10	0.03	0.04	0.52	0.45	6722
Federal Reserve Board	US	0.07	0.05	0.01	0.01	0.56	0.25	6720
University of Georgia	US	0.08	0.07	0.01	0.02	0.55	0.24	6551
Purdue University	US	0.14	0.12	0.02	0.03	0.54	0.28	6418
University of Southern California	US	0.16	0.14	0.05	0.04	0.46	0.21	6118
Northwestern University	US	0.09	0.07	0.02	0.02	0.53	0.27	5760
University of Cambridge	GB	0.10	0.09	0.02	0.02	0.53	0.24	5738
University of Wisconsin Madison	US	0.14	0.11	0.03	0.03	0.55	0.23	5722
University College London	GB	0.12	0.10	0.02	0.02	0.49	0.16	5451
Vanderbilt University	US	0.10	0.08	0.02	0.03	0.59	0.29	5366
Bates College	US	0.14	0.12	0.03	0.03	0.49	0.19	5194
University of Hong Kong	HK	0.09	0.07	0.02	0.02	0.51	0.32	5127
Princeton University	US	0.11	0.10	0.02	0.02	0.50	0.21	5029
University of California Irvine	US	0.12	0.09	0.03	0.03	0.55	0.27	4753
University of Pennsylvania	US	0.07	0.06	0.01	0.02	0.44	0.20	4662
Texas Tech University	US	0.14	0.11	0.03	0.05	0.55	0.32	4534
Cornell University	US	0.11	0.09	0.01	0.02	0.63	0.27	4510
Duke University	US	0.10	0.09	0.01	0.02	0.52	0.24	4307
UNC Chapel Hill	US	0.12	0.10	0.02	0.02	0.68	0.29	4281
Michigan State University	US	0.14	0.11	0.02	0.03	0.53	0.27	4177
California Institute of Technology	US	0.10	0.09	0.01	0.02	0.47	0.21	4085
Goethe Universitaet Frankfurt	DE	0.10	0.08	0.01	0.02	0.54	0.25	3994

Table 3 – continued from previous page

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
University of Virginia	US	0.20	0.18	0.03	0.03	0.64	0.23	3915
World Bank Group	US	0.07	0.05	0.01	0.02	0.62	0.34	3894
University of Wisconsin Milwaukee	US	0.11	0.09	0.03	0.02	0.72	0.27	3890
London School of Economics	GB	0.10	0.08	0.02	0.03	0.44	0.24	3826
Yale University	US	0.14	0.12	0.03	0.03	0.62	0.29	3749
University of British Columbia	CA	0.14	0.12	0.04	0.02	0.55	0.24	3716
Georgetown University	US	0.10	0.08	0.01	0.03	0.50	0.30	3677
Harvard University	US	0.10	0.09	0.01	0.02	0.49	0.20	3653
Australian National University	AU	0.11	0.10	0.02	0.03	0.51	0.25	3569
Hong Kong Polytechnic University	HK	0.07	0.04	0.03	0.02	0.62	0.30	3457
Wayne State University	US	0.06	0.05	0.01	0.01	0.45	0.17	3438
University of Illinois	US	0.12	0.09	0.02	0.03	0.48	0.20	3383
LMU Muenchen	DE	0.04	0.03	0.01	0.01	0.38	0.23	3345
University of Michigan	US	0.11	0.09	0.02	0.02	0.42	0.18	3197
Ohio State University	US	0.12	0.09	0.03	0.02	0.51	0.22	3147
University of Minnesota	US	0.10	0.08	0.02	0.03	0.61	0.27	3131
University of Tennessee	US	0.09	0.07	0.02	0.02	0.41	0.23	3073
SUNY Buffalo	US	0.08	0.06	0.01	0.02	0.47	0.18	3044
University of Dhaka	BD	0.27	0.25	0.09	0.08	0.65	0.25	3001
University of Toronto	CA	0.20	0.17	0.06	0.05	0.56	0.29	2942
Imperial College	GB	0.13	0.11	0.03	0.02	0.57	0.24	2924
Chinese University of Hong Kong	HK	0.07	0.04	0.02	0.02	0.61	0.25	2922
City University of Hong Kong	HK	0.10	0.08	0.02	0.02	0.57	0.26	2901
Universidade de Sao Paulo	BR	0.14	0.12	0.03	0.03	0.47	0.19	2874
UMichigan College of LS&A	US	0.16	0.15	0.02	0.02	0.54	0.18	2828
University of Miami	US	0.11	0.09	0.03	0.02	0.46	0.19	2814
Southern Methodist University	US	0.08	0.07	0.01	0.02	0.42	0.18	2800
Universitaet Mannheim	DE	0.11	0.09	0.02	0.01	0.44	0.19	2706
Chapman University	US	0.26	0.24	0.05	0.02	0.57	0.19	2649
Universitaet Zuerich	CH	0.14	0.12	0.03	0.02	0.62	0.26	2619
University of Texas at Austin	US	0.08	0.07	0.01	0.02	0.54	0.26	2548
University of Arizona	US	0.11	0.09	0.02	0.03	0.41	0.25	2517
ALU Freiburg	DE	0.09	0.07	0.02	0.02	0.67	0.31	2511
University of Nebraska-Lincoln	US	0.18	0.14	0.04	0.04	0.65	0.34	2487
University of Alberta	CA	0.09	0.07	0.01	0.02	0.53	0.29	2444
University of California San Diego	US	0.13	0.12	0.03	0.02	0.46	0.22	2431
University of Nottingham	GB	0.15	0.13	0.04	0.02	0.68	0.29	2387
University of Illinois at Chicago	US	0.15	0.13	0.03	0.03	0.49	0.20	2385
University of California Los Angeles	US	0.15	0.14	0.06	0.02	0.55	0.26	2368
University of Calgary	CA	0.13	0.10	0.02	0.03	0.46	0.19	2357
New York University	US	0.09	0.08	0.02	0.02	0.45	0.21	2280
Queen's University	CA	0.08	0.07	0.01	0.02	0.51	0.15	2223
University of Oregon	US	0.15	0.11	0.03	0.04	0.65	0.27	2207
University of California Davis	US	0.14	0.11	0.04	0.05	0.50	0.21	2173

Table 3 – continued from previous page

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
Georgia State University	US	0.05	0.04	0.01	0.01	0.43	0.28	2173
Indian School of Business	IN	0.12	0.09	0.04	0.04	0.40	0.21	2125
Humboldt-Universitaet zu Berlin	DE	0.09	0.07	0.02	0.01	0.60	0.33	2105
Virginia Wesleyan University	US	0.25	0.19	0.06	0.08	0.66	0.26	2046
University of Kentucky	US	0.09	0.07	0.01	0.02	0.60	0.22	2042
University of South Florida	US	0.10	0.08	0.02	0.03	0.44	0.29	2005
Lingnan University	HK	0.23	0.18	0.06	0.10	0.49	0.28	1983
University of Texas at Dallas	US	0.19	0.15	0.04	0.05	0.61	0.30	1983
University of Western Ontario	CA	0.11	0.09	0.02	0.03	0.52	0.24	1972
SUNY New Paltz	US	0.14	0.12	0.03	0.02	0.36	0.24	1942
Clemson University	US	0.06	0.05	0.01	0.01	0.44	0.23	1908
Universitaet Heidelberg	DE	0.13	0.10	0.03	0.03	0.51	0.25	1883
Universidad Pablo de Olavide	ES	0.05	0.05	0.01	0.00	0.52	0.40	1873
Emory University	US	0.13	0.11	0.03	0.02	0.60	0.26	1837
North Dakota State University	US	0.09	0.08	0.01	0.01	0.68	0.27	1806
Rice University	US	0.25	0.22	0.07	0.04	0.69	0.29	1804
Korea Advanced Institute	KR	0.05	0.04	0.01	0.01	0.45	0.27	1766
University of Lancaster	GB	0.07	0.06	0.01	0.01	0.62	0.37	1763
Brown University	US	0.09	0.06	0.02	0.03	0.54	0.24	1756
Washington University	US	0.10	0.09	0.01	0.02	0.65	0.29	1725
University of Edinburgh	GB	0.14	0.12	0.03	0.02	0.62	0.26	1719
Kansas State University	US	0.12	0.09	0.01	0.03	0.65	0.26	1716
Montclair State University	US	0.13	0.10	0.02	0.04	0.47	0.30	1715
RFWU Bonn	DE	0.14	0.11	0.04	0.04	0.57	0.27	1659
Monash University	AU	0.12	0.09	0.03	0.02	0.53	0.23	1647
Universitaet Kiel	DE	0.05	0.04	0.01	0.01	0.48	0.25	1623
University of California Santa Cruz	US	0.12	0.10	0.04	0.02	0.44	0.20	1613
Seoul National University	KR	0.09	0.07	0.01	0.02	0.67	0.21	1598
Loyola Marymount University	US	0.03	0.03	0.00	0.00	0.79	0.21	1598
University of Waterloo	CA	0.12	0.07	0.03	0.07	0.34	0.26	1538
Charles University	CZ	0.07	0.06	0.01	0.01	0.51	0.23	1536
Temple University	US	0.10	0.09	0.01	0.02	0.46	0.20	1506
University of California Merced	US	0.02	0.02	0.00	0.01	0.41	0.23	1496
International Monetary Fund	US	0.14	0.12	0.03	0.03	0.60	0.30	1495
College of William and Mary	US	0.13	0.12	0.02	0.02	0.70	0.28	1494
University of Montreal	CA	0.06	0.04	0.01	0.02	0.43	0.20	1490
Indiana University	US	0.10	0.07	0.02	0.03	0.60	0.31	1426
Cornell Weill Medical College	US	0.11	0.08	0.02	0.04	0.53	0.23	1425
University of California	US	0.11	0.08	0.02	0.03	0.57	0.34	1393
Bogazici University	TR	0.10	0.08	0.02	0.03	0.50	0.35	1389
University of Otago	NZ	0.09	0.06	0.01	0.04	0.45	0.27	1378
Kyungpook National University	KR	0.04	0.03	0.00	0.01	0.53	0.14	1374
Aalto University	FI	0.11	0.09	0.02	0.03	0.68	0.31	1365
CSU de Catalunya	ES	0.13	0.11	0.02	0.02	0.65	0.33	1341

Table 3 – continued from previous page

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
UMichigan Engineering	US	0.09	0.07	0.02	0.03	0.47	0.23	1322
University of California Riverside	US	0.10	0.08	0.01	0.03	0.50	0.21	1318
Brock University	CA	0.03	0.03	0.01	0.00	0.41	0.23	1305
NYU Abu Dhabi	AE	0.17	0.12	0.02	0.07	0.70	0.30	1293
University of Exeter	GB	0.16	0.13	0.04	0.03	0.65	0.29	1272
George Mason University	US	0.08	0.07	0.01	0.02	0.50	0.21	1257
IIM Ahmedabad	IN	0.21	0.17	0.05	0.03	0.75	0.25	1254
SUNY Stony Brook	US	0.14	0.12	0.03	0.02	0.69	0.27	1239
Carnegie Mellon University	US	0.10	0.08	0.02	0.02	0.46	0.22	1222
Louisiana State University	US	0.09	0.07	0.02	0.03	0.54	0.25	1214
Suny College at Fredonia	US	0.07	0.05	0.01	0.02	0.69	0.29	1167
Tulane University	US	0.08	0.06	0.01	0.02	0.55	0.27	1148
Suffolk Community College	US	0.14	0.11	0.03	0.03	0.69	0.30	1140
University of Missouri-Columbia	US	0.08	0.08	0.02	0.01	0.57	0.26	1138
UMichigan Network	US	0.09	0.06	0.02	0.04	0.49	0.25	1121
Mississippi State University	US	0.16	0.09	0.04	0.09	0.57	0.31	1113
Baylor University	US	0.07	0.06	0.00	0.01	0.57	0.19	1108
UMichigan School of Information	US	0.10	0.08	0.03	0.02	0.48	0.20	1099
Dartmouth College	US	0.07	0.06	0.01	0.02	0.48	0.16	1077
Washington State University	US	0.10	0.08	0.01	0.03	0.71	0.26	1068
University of Essex	GB	0.10	0.08	0.04	0.02	0.39	0.20	1040
West Texas A&M University	US	0.08	0.07	0.01	0.01	0.67	0.30	1017
NUS	SG	0.09	0.06	0.01	0.02	0.61	0.34	1011
Florida International University	US	0.07	0.05	0.02	0.02	0.66	0.29	999
Virginia Commonwealth University	US	0.16	0.12	0.03	0.07	0.63	0.29	996
University of Vienna	AT	0.14	0.12	0.02	0.02	0.60	0.31	955
University of Glasgow	GB	0.06	0.05	0.01	0.01	0.52	0.22	951
Middle Tennessee State University	US	0.09	0.07	0.01	0.02	0.58	0.25	946
University of Warwick	GB	0.07	0.06	0.01	0.01	0.58	0.25	937
Queen Mary University of London	GB	0.12	0.10	0.01	0.03	0.52	0.21	935
University of Connecticut	US	0.07	0.07	0.01	0.01	0.45	0.20	933
University of Texas at Arlington	US	0.06	0.04	0.01	0.02	0.61	0.22	912
University of New South Wales	AU	0.15	0.11	0.07	0.04	0.51	0.27	885
University of Alabama	US	0.12	0.11	0.02	0.02	0.56	0.31	848
Singapore Management University	SG	0.12	0.10	0.02	0.01	0.65	0.31	827
London Business School	GB	0.12	0.07	0.04	0.04	0.46	0.26	825
Rise at State College	US	0.15	0.10	0.03	0.04	0.62	0.38	807
University of Bristol	GB	0.14	0.12	0.04	0.03	0.55	0.28	806
Auburn University	US	0.12	0.10	0.01	0.02	0.73	0.20	791
Our Lady of the Lake University	US	0.12	0.08	0.02	0.04	0.68	0.32	785
University of Colorado	US	0.08	0.06	0.01	0.02	0.50	0.22	772
University of New Mexico	US	0.06	0.05	0.01	0.01	0.60	0.31	762
Boston University	US	0.10	0.07	0.01	0.04	0.59	0.32	740
Binghamton University	US	0.09	0.08	0.00	0.01	0.69	0.30	729

Table 3 – continued from previous page

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
UNC Greensboro	US	0.10	0.08	0.01	0.02	0.48	0.24	728
McMaster University	CA	0.15	0.14	0.04	0.02	0.54	0.31	722
Sogang University	KR	0.02	0.02	0.01	0.01	0.41	0.18	715
University of Arkansas	US	0.20	0.14	0.06	0.06	0.75	0.21	713
University of Houston	US	0.14	0.12	0.04	0.03	0.47	0.18	710
Universidad de Guadalajara	MX	0.15	0.13	0.02	0.03	0.49	0.26	689
University of Hawaii	US	0.37	0.36	0.09	0.20	0.73	0.26	670
Universitaet St. Gallen	CH	0.06	0.04	0.01	0.01	0.45	0.23	669
Northern Illinois University	US	0.20	0.16	0.04	0.06	0.47	0.21	668
Iowa State University	US	0.02	0.02	0.00	0.00	0.33	0.28	639
Otto von Guericke Universitaet	DE	0.15	0.13	0.03	0.02	0.54	0.24	620
American University of Iraq-Sulaimani	IQ	0.22	0.19	0.04	0.03	0.70	0.30	619
Indiana University Health	US	0.05	0.05	0.01	0.01	0.38	0.19	619
University of Helsinki	FI	0.05	0.05	0.00	0.01	0.42	0.23	615
Oberlin College	US	0.06	0.05	0.01	0.01	0.47	0.19	588
University of Central Missouri	US	0.17	0.14	0.02	0.04	0.61	0.28	573
Westfaelische Wilhelms-Universitaet	DE	0.05	0.05	0.01	0.01	0.50	0.15	531
University of Florida	US	0.18	0.17	0.04	0.03	0.49	0.17	527
Johns Hopkins University	US	0.07	0.04	0.02	0.02	0.48	0.20	526
University of Wisconsin Stout	US	0.15	0.14	0.03	0.04	0.50	0.19	520
Urban Institute	US	0.07	0.06	0.01	0.01	0.53	0.24	519
University of York	GB	0.16	0.13	0.02	0.03	0.69	0.30	518
University of St. Andrews	GB	0.20	0.17	0.06	0.04	0.58	0.32	507
Lehigh University	US	0.07	0.06	0.01	0.01	0.67	0.31	497
North Carolina State University	US	0.10	0.08	0.01	0.01	0.63	0.32	495
Drexel University	US	0.02	0.02	0.00	0.01	0.38	0.20	494
Georgia Institute of Technology	US	0.08	0.06	0.01	0.01	0.48	0.25	492
Suffolk University	US	0.05	0.03	0.00	0.01	0.65	0.29	485
Claremont University Consortium	US	0.16	0.15	0.03	0.03	0.53	0.20	482
St. John's College	US	0.12	0.09	0.04	0.00	0.80	0.20	480
Tufts University	US	0.09	0.07	0.01	0.03	0.61	0.37	474
University of South Carolina	US	0.04	0.02	0.01	0.01	0.69	0.31	472
UMichigan School of Business	US	0.06	0.05	0.00	0.00	0.55	0.27	471
Rhode Island College	US	0.06	0.05	0.01	0.00	0.56	0.28	471
University of Utah	US	0.10	0.09	0.03	0.03	0.41	0.27	467
University of Central Florida	US	0.02	0.02	0.00	0.00	0.43	0.17	459
University of Arkansas Little Rock	US	0.25	0.16	0.03	0.11	0.68	0.31	446
University of Maryland Baltimore	US	0.07	0.07	0.01	0.02	0.50	0.23	445
Baruch College	US	0.04	0.03	0.00	0.01	0.52	0.30	439
University of Oklahoma	US	0.09	0.08	0.01	0.02	0.69	0.31	426
Arizona State University	US	0.06	0.05	0.00	0.00	0.64	0.36	426
Oklahoma State University	US	0.06	0.05	0.01	0.01	0.45	0.29	424
Syracuse University	US	0.09	0.07	0.02	0.02	0.41	0.23	422
University of California Santa Barbara	US	0.09	0.07	0.02	0.02	0.56	0.28	404

Table 3 – continued from previous page

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
Vienna University of Economics	AT	0.10	0.08	0.01	0.03	0.69	0.30	394
ETH Zurich	CH	0.08	0.06	0.03	0.01	0.66	0.34	393
Pontificia Universidad Javeriana	CO	0.09	0.09	0.01	0.02	0.57	0.24	392
CSU Office of the Chancellor	US	0.06	0.06	0.01	0.01	0.46	0.25	386
Rutgers	US	0.13	0.11	0.03	0.05	0.46	0.19	383
Florida State University	US	0.07	0.05	0.01	0.02	0.59	0.22	366
Bilkent University	TR	0.18	0.15	0.03	0.02	0.54	0.26	364
International Institute for ASA	AT	0.23	0.18	0.04	0.04	0.68	0.32	363
University of Delaware	US	0.09	0.04	0.01	0.06	0.63	0.33	361
Hunter College	US	0.10	0.07	0.03	0.05	0.34	0.18	354
University of Economics Prague	CZ	0.06	0.06	0.01	0.01	0.51	0.49	342
University of Cincinnati	US	0.14	0.11	0.03	0.03	0.66	0.31	335
University of Durham	GB	0.08	0.06	0.01	0.03	0.42	0.21	332
Texas A&M	US	0.13	0.11	0.02	0.04	0.55	0.27	330
Simon Fraser University	CA	0.12	0.10	0.03	0.04	0.54	0.22	330
Universitaet Bremen	DE	0.10	0.07	0.02	0.03	0.49	0.20	323
University of Surrey	GB	0.20	0.18	0.05	0.02	0.66	0.23	323
Queens University Belfast	GB	0.14	0.13	0.01	0.02	0.64	0.24	321
Acadia University	CA	0.08	0.08	0.01	0.01	0.65	0.23	319
University of Lausanne	CH	0.09	0.07	0.02	0.03	0.61	0.28	317
University of Iowa	US	0.06	0.05	0.01	0.02	0.50	0.19	316
Thompson Rivers University	CA	0.06	0.05	0.01	0.02	0.44	0.17	314
San Diego State University	US	0.04	0.03	0.00	0.01	0.69	0.31	300
University of Massachusetts	US	0.10	0.09	0.01	0.02	0.64	0.25	299
University of Albany	US	0.08	0.07	0.00	0.02	0.71	0.29	291
UMichigan Life Sciences Institute	US	0.11	0.07	0.01	0.05	0.55	0.25	287
Washington School Information	US	0.12	0.08	0.04	0.04	0.38	0.28	282
East China UST	CN	0.04	0.02	0.03	0.00	0.41	0.17	280
West Virginia University	US	0.09	0.09	0.01	0.00	0.64	0.35	280
Universitaet Bern	CH	0.05	0.05	0.00	0.00	0.50	0.26	275
University of Leeds	GB	0.15	0.12	0.03	0.01	0.67	0.32	275
University of Pittsburgh	US	0.03	0.03	0.01	0.00	0.62	0.26	271
Brandeis University	US	0.09	0.09	0.02	0.01	0.45	0.20	269
Villanova University	US	0.12	0.10	0.03	0.02	0.62	0.33	269
PU Marburg	DE	0.07	0.05	0.01	0.02	0.51	0.22	267
Wright State University	US	0.18	0.15	0.05	0.07	0.48	0.24	265
IFO Institut	DE	0.10	0.10	0.01	0.00	0.66	0.28	262
Kennesaw State University	US	0.14	0.10	0.05	0.01	0.55	0.25	259
Whitworth University	US	0.19	0.12	0.04	0.06	0.69	0.31	257
Virginia Polytechnic Institute	US	0.16	0.14	0.02	0.04	0.62	0.38	251
University of Ghent	BE	0.06	0.03	0.00	0.03	0.49	0.38	251
Sam Houston State University	US	0.27	0.25	0.05	0.04	0.77	0.23	249
Miami University	US	0.05	0.05	0.00	0.00	0.59	0.33	247
MIT	US	0.12	0.10	0.03	0.02	0.58	0.37	245

Table 3 – continued from previous page

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
Texas A&M San Antonio	US	0.11	0.09	0.04	0.03	0.69	0.30	242
University of East Anglia	GB	0.05	0.05	0.00	0.02	0.42	0.24	239
Nova University	US	0.16	0.12	0.06	0.03	0.49	0.14	232
Flinders University	AU	0.03	0.03	0.01	0.01	0.42	0.24	231
Lake Forest College	US	0.12	0.09	0.01	0.04	0.48	0.25	221
University of Innsbruck	AT	0.16	0.13	0.04	0.03	0.62	0.35	216
Valdosta State University	US	0.12	0.10	0.04	0.07	0.40	0.19	212
UMichigan Dearborn	US	0.05	0.04	0.01	0.01	0.65	0.35	212
LaTrobe University	AU	0.17	0.14	0.06	0.08	0.48	0.38	208
Stockholm University	SE	0.06	0.05	0.00	0.01	0.42	0.19	207
Ball State University	US	0.23	0.23	0.07	0.01	0.47	0.14	206
University of North Florida	US	0.03	0.02	0.01	0.01	0.47	0.24	202
UMichigan Academic Affairs	US	0.09	0.08	0.03	0.03	0.46	0.18	197
Kenya School of Monetary Studies	KE	0.03	0.03	0.00	0.01	0.33	0.26	194
University of Melbourne	AU	0.06	0.05	0.01	0.01	0.37	0.22	194
University de Los Andes	CO	0.11	0.07	0.02	0.03	0.40	0.21	184
American University	US	0.07	0.04	0.01	0.03	0.54	0.29	184
Grand Valley State University	US	0.04	0.03	0.00	0.01	0.29	0.26	184
Lund University	SE	0.12	0.09	0.01	0.04	0.70	0.30	184
University of Kent	GB	0.04	0.03	0.01	0.01	0.50	0.43	181
Central Michigan University	US	0.11	0.10	0.02	0.03	0.65	0.25	179
Norwegian School of Management	NO	0.15	0.11	0.03	0.01	0.70	0.27	178
Universidad Carlos III de Madrid	ES	0.10	0.06	0.05	0.00	0.51	0.24	176
McGill University	CA	0.07	0.06	0.01	0.01	0.49	0.31	175
Universidad de Piura	PE	0.26	0.24	0.06	0.05	0.75	0.25	175
Sacred Heart University	US	0.11	0.10	0.01	0.01	0.40	0.15	174
Morgan State University	US	0.04	0.03	0.00	0.01	0.64	0.36	171
University of Manchester	GB	0.05	0.04	0.02	0.02	0.44	0.27	164
University of Wyoming	US	0.06	0.06	0.00	0.01	0.39	0.21	163
Institute for Advanced Study	US	0.05	0.04	0.00	0.01	0.38	0.11	161
Hong Kong Baptist University	HK	0.03	0.00	0.02	0.01	0.48	0.45	153
Universidad de Alicante	ES	0.15	0.11	0.01	0.04	0.35	0.22	150
University of Northumbria Newcastle	GB	0.14	0.13	0.02	0.02	0.51	0.18	148
Carleton University	CA	0.10	0.07	0.01	0.03	0.65	0.33	147
University of Prince Edward Island	CA	0.07	0.06	0.02	0.03	0.63	0.32	145
George Washington University	US	0.01	0.00	0.00	0.01	0.45	0.16	143
Swarthmore College	US	0.07	0.05	0.01	0.02	0.59	0.23	139
Clark University	US	0.02	0.02	0.00	0.00	0.49	0.30	135
IIM Kashipur	IN	0.05	0.04	0.02	0.00	0.53	0.44	131
University of Texas at San Antonio	US	0.08	0.08	0.01	0.00	0.48	0.15	130
Berry College	US	0.03	0.03	0.00	0.00	0.57	0.24	127
Purdue University Fort Wayne	US	0.20	0.15	0.04	0.09	0.53	0.46	127
Northern Arizona University	US	0.25	0.16	0.02	0.11	0.67	0.31	127
Hawaii Pacific University	US	0.17	0.17	0.01	0.02	0.70	0.30	126

Table 3 – continued from previous page

ISP and Country Code		Combined	Separate Labels			Sentiment		Count
			Toxicity	Hate Speech	Misogyny	+	-	
Boston College	US	0.14	0.10	0.02	0.03	0.65	0.34	125
Universitaet Hamburg	DE	0.03	0.02	0.01	0.01	0.48	0.17	124
International Christian University	JP	0.08	0.07	0.01	0.01	0.46	0.11	122
Austin College	US	0.04	0.03	0.01	0.01	0.49	0.08	118
Ithaca College	US	0.11	0.07	0.02	0.04	0.40	0.17	115
Southern Utah University	US	0.07	0.07	0.00	0.00	0.62	0.27	114
University of the Witwatersrand	ZA	0.07	0.07	0.02	0.00	0.44	0.14	113
University of Auckland	NZ	0.12	0.11	0.02	0.03	0.34	0.27	112
Augustana College	US	0.04	0.04	0.00	0.01	0.52	0.25	112
Brandenburgische TU	DE	0.16	0.11	0.04	0.05	0.45	0.21	110
Western Illinois University	US	0.18	0.15	0.02	0.05	0.42	0.28	110
Fordham University	US	0.04	0.04	0.00	0.00	0.44	0.23	108
Dalhousie University	CA	0.07	0.05	0.01	0.03	0.59	0.15	107
University of Newcastle upon Tyne	GB	0.25	0.20	0.06	0.03	0.73	0.19	105
University of Kansas	US	0.06	0.05	0.01	0.02	0.35	0.22	105
Guangzhou University	CN	0.03	0.02	0.01	0.01	0.43	0.11	103
Amherst College	US	0.09	0.09	0.00	0.00	0.46	0.19	102
Colby College	US	0.05	0.04	0.00	0.02	0.39	0.13	102
Melbourne Institute of Technology	AU	0.01	0.00	0.01	0.00	0.64	0.36	101