Anonymity and Identity Online*

Florian Ederer† Paul Goldsmith-Pinkham‡ Kyle Jensen§

July 17, 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 much 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 match these 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 speech—across sub-forums, geographies, institutions, and contributors.

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

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

*We are grateful to DAC, Forrest Crawford, Rachael Meager, Barry Nalebuff, and Julia Simon-Kerr for helpful comments. The research described in this manuscript was determined to be exempt by the Yale University IRB, protocol 2000034072.

†Boston University, 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 17 May 2023 after the first publication of the abstract of this paper on 16 May 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 which provides a place to share information relevant to economics, most notably about the job market for PhD economists. However, the discussion board is active year-round and EJMR users post anonymously about economics-related or miscellaneous issues. EJMR is very popular: SimilarWeb estimates that EJMR receives 2.5 million visits per month with an average of 6.45 pages viewed per visit. In comparison, the same figures for the NBER and AEA competitors are 1.1 million and 991,000 visits and 2.09 and 2.76 pages per visit, respectively. Beyond its informational role, EJMR is also "a breeding ground for personal attacks of an abusive kind" (Blanchard, 2017) and features widespread sexist speech first documented in a systematic manner by Wu (2018) and Wu (2020).

The use of anonymity on the site can lead to heated discussions, with users frequently making controversial and inflammatory statements, including personal attacks, harassment, and discrimination. The informal and unverified nature of the information shared on EJMR is also controversial. Information posted on the site is not subject to fact-checking or verification, and there is a risk of misinformation being spread. This can lead to false rumors about job opportunities, candidates, and hiring processes. Furthermore, EJMR perpetuates and amplifies existing inequalities in the economics profession, such as gender and race-based discrimination (Bayer and Rouse, 2016; Antecol et al., 2018; Lundberg and Stearns, 2019; Dupas et al., 2021; Hengel, 2022). Despite automatic deletion of much offensive speech and heavy moderation there are numerous instances of sexist, racist, or otherwise discriminatory comments on the site. Sometimes these comments are general in nature and attack broad groups (e.g., "the whole point of women is to get railed and make babies" or "The biggest enemies of America are: Blks" or "And America lost its war against blks. [...] At least until we resolve to final solution." or "University of Stupid Chinese" or "The average woman has a 15% smaller brain than the average man" or "the fastest route to a qje is to grift and be black"). Other times they target particular individuals (e.g., "Should Jennifer Doleac be executed for her anti-Chinese hatred?" or "Anya Samek [...] took advantage of her initial postdoc position organizing conferences with JL and handling requests for grant proposals to steal ideas" or "Are Vrinda and Hampole in a secret same-sekhs love-hayte relationship?").

In response, more than 1,000 signers (https://www.iaffe.org/petition-aea-ejmr/) urged the American Economic Association to create a moderated, well-functioning site to

provide up-to-the-minute job market information. The AEA CSWEP Board (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." However, the forum EconSpark and the information website EconTrack, both created by the AEA, were largely failures and languish without much use while EJMR's user base remains large (see Figure 6).

EJMR is sometimes dismissed as not being representative of the economics profession, including claims that the most frequent users on the platform are not actually economists. However, our analysis reveals that the users who post on EJMR are predominantly economists, including those working in the upper echelons of academia, government, and the private sector.

In this paper, we identify the scheme used to assign usernames for each post written by an anonymous user on EJMR. We show how the statistical properties of that algorithm do not anonymize 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 roughly narrows the set of possible IP addresses from 2^{32} to 2^{16}. Second, we measure which IP addresses occur particularly often in a narrow 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 is very conservative and 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}. For example, even though there are 7,098,111 posts on EJMR, we assign exactly *zero* posts to the large set of bogon IP addresses. Despite this very 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. Based on the geographic content of the IP addresses we identify and the origin of the associated internet service providers, we show that posting on EJMR is pervasive and very common in academia. Over 10% of posts originate from universities 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. However, the vast majority of EJMR posts comes from residential IP addresses located in the United States and in particular in cities

with elite universities.

Like other online platforms posts on EJMR are very concentrated across posters. A mere 5% of IP addresses generate over 50% of posts and 20% generate more than 80% of posts and this concentration is even more pronounced for toxic posts.

However, there is also a significant share of other countries including Canada, the United Kingdom, Hong Kong, Australia, Germany, Italy, and France. EJMR users tend to post during work hours and in the evening of their respective time zones. Although EJMR has been a popular website for economists since its inception, posting and engagement on the site has surged since the start of the COVID-19 pandemic, especially in the United States.

We show that toxic speech is widespread on EJMR with more than 10% of posts classified as toxic. Toxic posts are more likely to originate from residential IP addresses than from IP addresses associated with universities, government agencies, and private sector institutions.

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 describes 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."[2] That IP address—twelve years ago—was almost certainly 188.220.40.122.[3]

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 identify nearly 100% of the IP addresses that were consistently active on the site and, in

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

[3] 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).

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. Though, 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, corporations, 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 has the same IP address for years potentially. 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 whilst 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.

The above should indicate that *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*.[4]

Having covered IP addresses, a short description of EJMR's organization is necessary.

[4]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.

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 that made the post.

As we described previously, most EJMR users do not log into the site using a self-selected username like users on a typical bbPress site would. 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 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.

We sought to understand how usernames are generated on the site. 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, 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."[5] 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

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

most practical uses of hash functions. As the EJMR notice 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 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. u and t we observed 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.[6] Our search was short. o, we set to null, presuming there was no salt. S we guessed was

[6] In total, we posted five times on EJMR: three times to verify the hashing scheme and two times to produce a brief video to document how the hashing scheme worked.

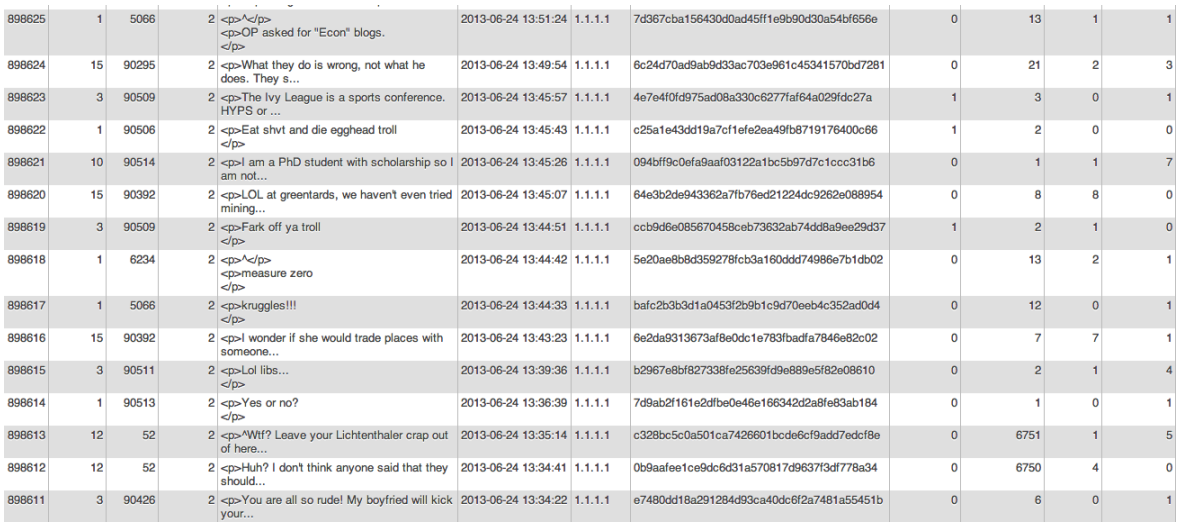
either the concatenation of t and a as strings, or a and t. M we guessed was a function that merely returned a substring of the hash. The hash function $\mathcal{H}$ we guessed was a common hash function such as MD5, SHA-1/224/256/384/512, or CRC32.

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 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 neigh 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 username. This video can be viewed at https://youtu.be/0dGyXlBMWWo. 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[7] 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

[7] 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."

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 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[8] 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

[8] https://web.archive.org/web/20150715234225/https://www.econjobrumors.com/topic/why-is-measure-theory-considered-useful-for-theory-and-finance#post-898618

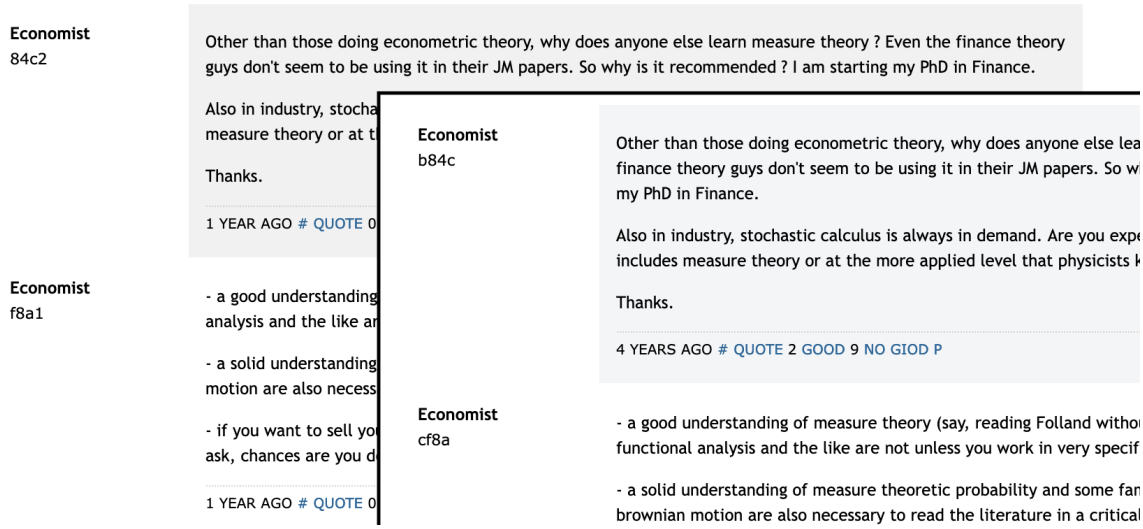


Figure 2: Screenshot of an EJMR topic before and after July 8, 2013, from the WayBack Machine. The leftmost posts shows usernames created using positions 10, 11, 12, and 13 in the SHA-1 hash of the topic id, IP address combination. The inset on the right shows the same posts after July 8, 2013 where 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.

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, it is helpful to offer some background on the SHA-1 hash. The SHA-1 hash was created by the U.S. 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: 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 choose a topic ID at random and computed the EJMR username for all IPv4 addresses. Second, we choose a random IPv4 address and compute the EJMR username for all extant EJMR topic IDs. In both cases we find the SHA-1's uniformity to be preserved over the two-bytes used for the EJMR username. Or, more formally, we cannot 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.[9]

2.3 Probabilistic Identification of IP Addresses

Recall that there are 2^{32} possible 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.[10]

Although this significantly narrows the set of possible IP addresses, it does not allow us to exactly identify 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 identifying 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.[11] 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 estimate π_a for all a.

[9] This code will be made available when our manuscript is made public.

[10] 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 $2^{16} + 1$.

[11] 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.

Moreover, we would like to estimate, for a given username, the probability that this 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, $\sum_{u \in U_t} 1(a \in A_{(t,u)}) = 0$, 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 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. 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 allow for multiple topics m, and consider the data generating process in this setting. 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. To simplify notation, let $y_t = \sum_{u \in U_t} 1(a \in A_{(t,u)})$ denote the binary variable of whether we observe IP a in the set of possible IPs for topic t. Then, for an $m \times 1$ vector $\boldsymbol{y}$ of

the observed y_t, we have

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

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

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

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.[12]

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 (e.g., Hochberg and Tamhane (1987); Benjamini and Hochberg (1995)) because for each post we test whether any of the $\tilde{6}5$k 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 $N\pi_0$ 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).

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

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 both intuitive and tractable for us.

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—and lacking a proper Bayesian model—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.[13] 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 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 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-8 prior to July 13, 2013 and position-10 thereafter until May 17, 2023).

[13] 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 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 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.[14] 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.[15] 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

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

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

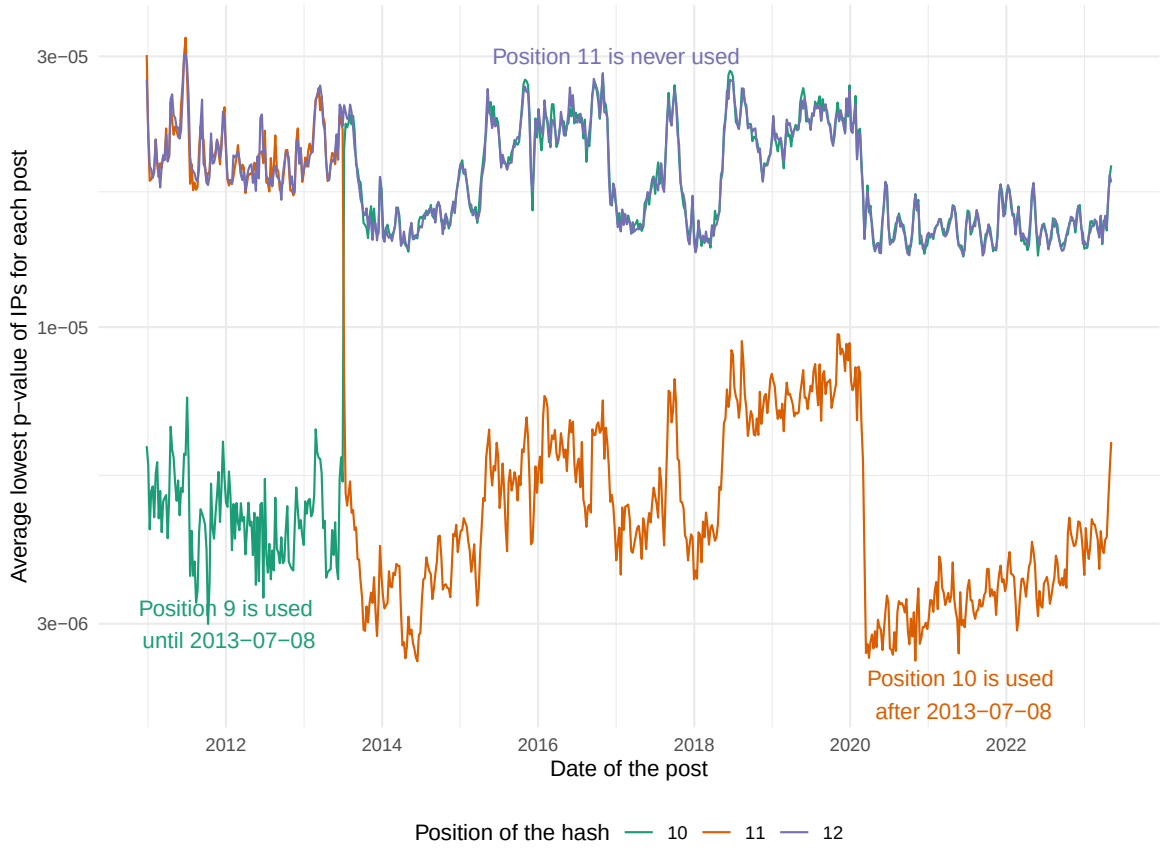


Figure 3: Average minimum p-value of posts in a given week for different hash positions over time. We employ the following two-step procedure. First, we find the IP address that has the lowest p-value for a given post and refer to its p-value as the minimum p-value of a post. Second, we calculate the 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.

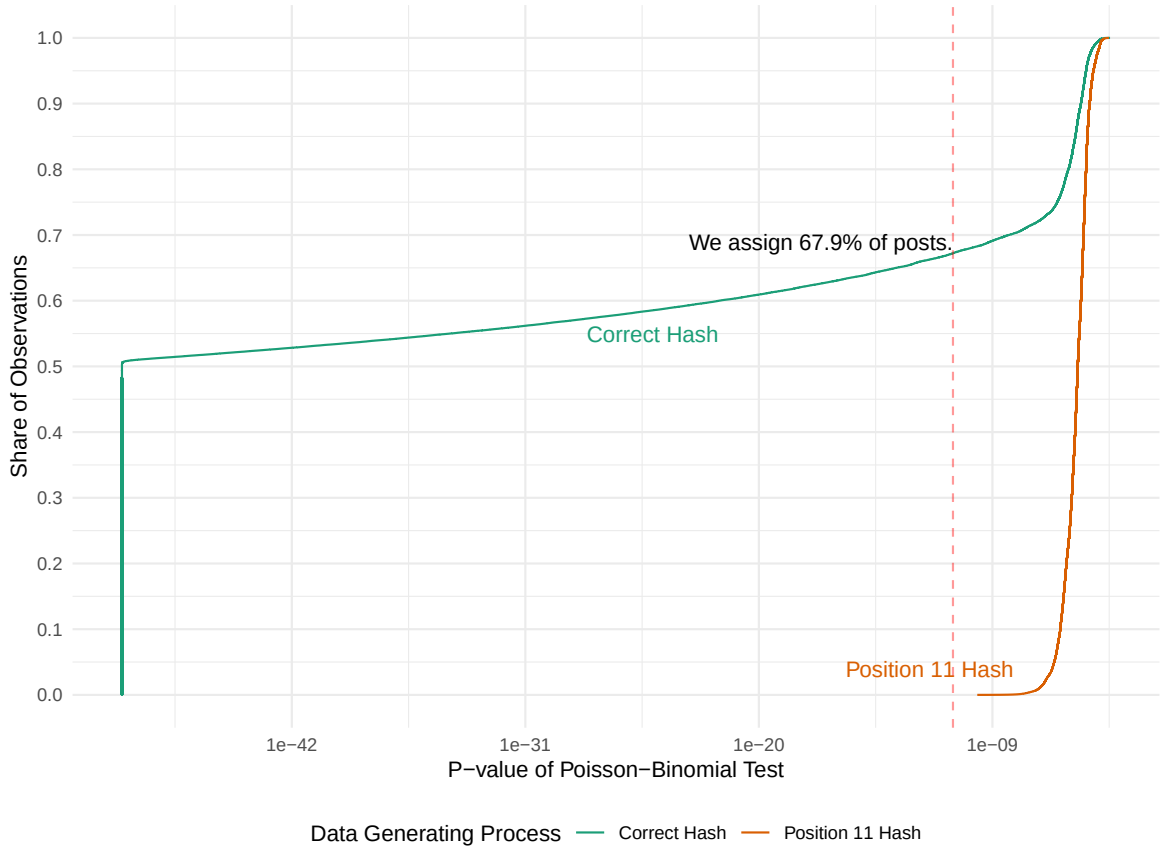


Figure 4: Cumulative distribution functions of the minimum p-value of posts for different hash positions. The orange line plots the CDF of the minimum p-values for posts calculated based on the incorrect position-11 hash. The green line shows the CDF under the correct hash (position-9 before July 8, 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, 66.1% of posts have minimum p-values below p^* and for more than 50% of posts the minimum p-values are smaller than 10^{-50}.

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). 66.1% 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 have 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 type of error, imagine the following situation. A one-time EJMR user posts to some topic and receives username ab34. It happens that a highly active IP address would also receive that same username. That is, 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 that 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. And, 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. 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 appeared to be state-of-the art and that also appeared to perform well on EJMR content, based on our informal inspection. 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 Toxi-Gen Roberta (Hartvigsen et al., 2022). This is a checkpoint of the Roberta model (Liu et al., 2019) fined-tuned for toxicity detection. We also re-created all the word-count measures used in Wu (2020).

2.6 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 networks of which they make use. Recent research suggests that the average consumer IP address in the US 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).

2.7 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 giving us exact time stamps for a total of 3,689,727 posts. For the remaining 3,408,384 posts, we assign 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 are 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

In this section we describe the 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. From these data we recover 47,630 distinct IP addresses.

3.2 Time Patterns

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.

Figure 6 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 6, 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 Figure 7 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.

The 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 8. However, unlike the United States the posting intensity in these countries mostly returned to pre-pandemic levels. Posts for which

Figure 5: Number of EJMR posts over time. 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.

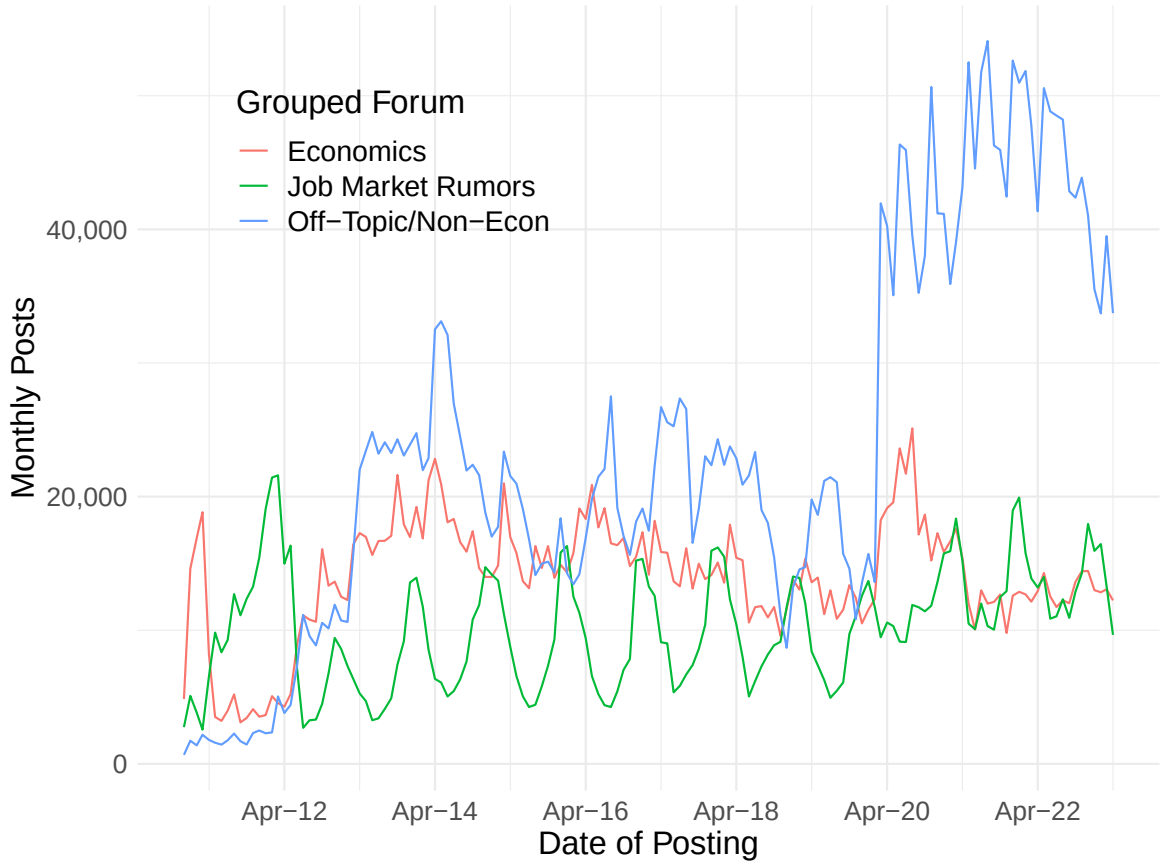


Figure 6: Distribution of posts by grouped forums over time. The figure shows the number of monthly posts in three groups of forums. Economics contains all forums with general economics discussions. Job Market Rumors contains all forums related to the academic job market including both junior and senior hiring. Off-Topic/Non-Econ contains all other forums.

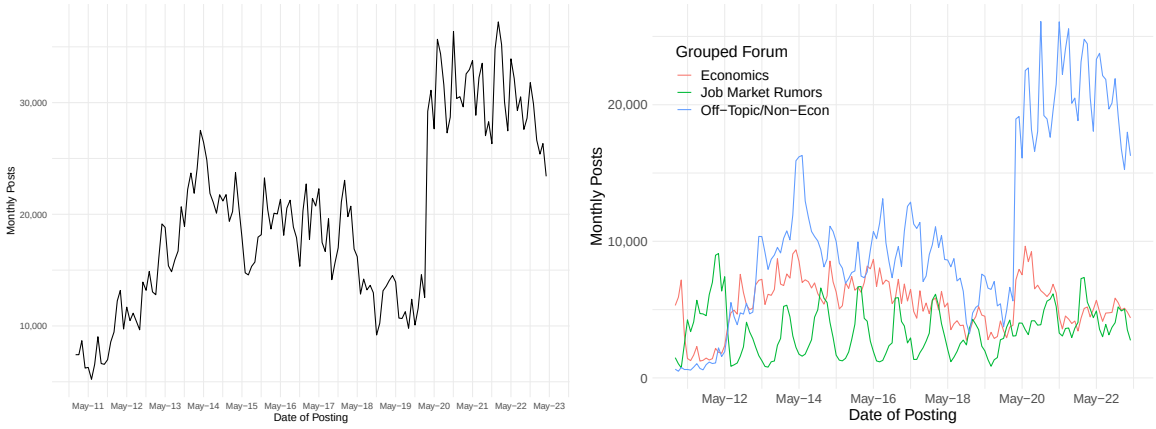


Figure 7: Distribution of aggregate posts and by grouped forums over time for the United States.

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.

EJMR users tend to primarily post during US work hours and to a lesser extent in the evening. Figure 9 shows total number of posts per minute by year from 2016 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 10 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, most of the posts originate from IP addresses located in the United States.

3.3 Geographical Distribution

The vast majority of EJMR posts originates from IP addresses located in the United States. As noted in Section 2.2 we are able to assign 66.1% of posts to IP addresses.

Figure 11 shows that we are able to assign 67.8% of 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 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 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 12 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).

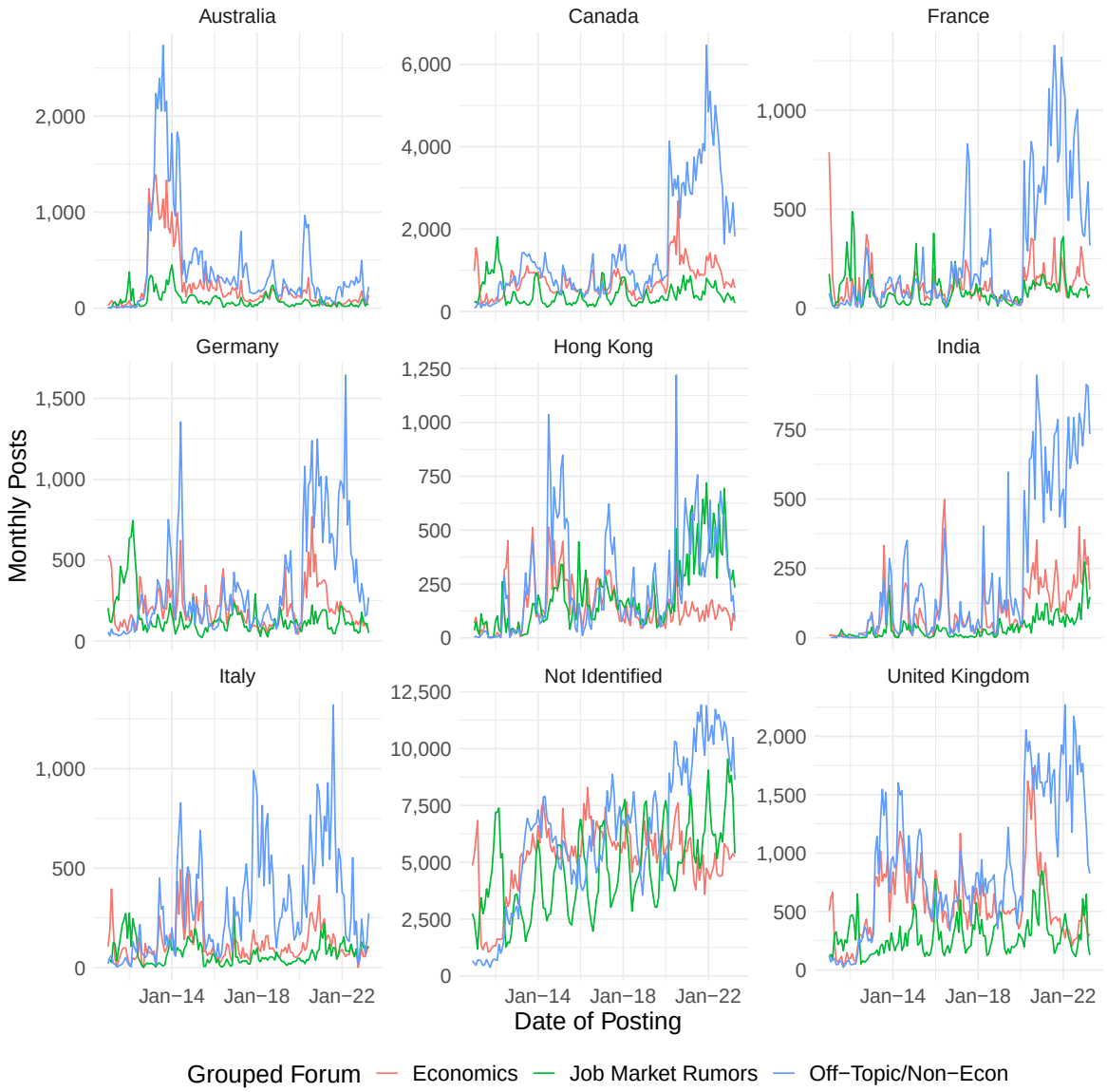


Figure 8: Monthly posts by grouped forums and by country (excluding US). The figure shows the distribution of monthly posts across the three large forum groups for the eight countries with the largest number of posts after the US and for those posts for which we do not assign IP addresses.

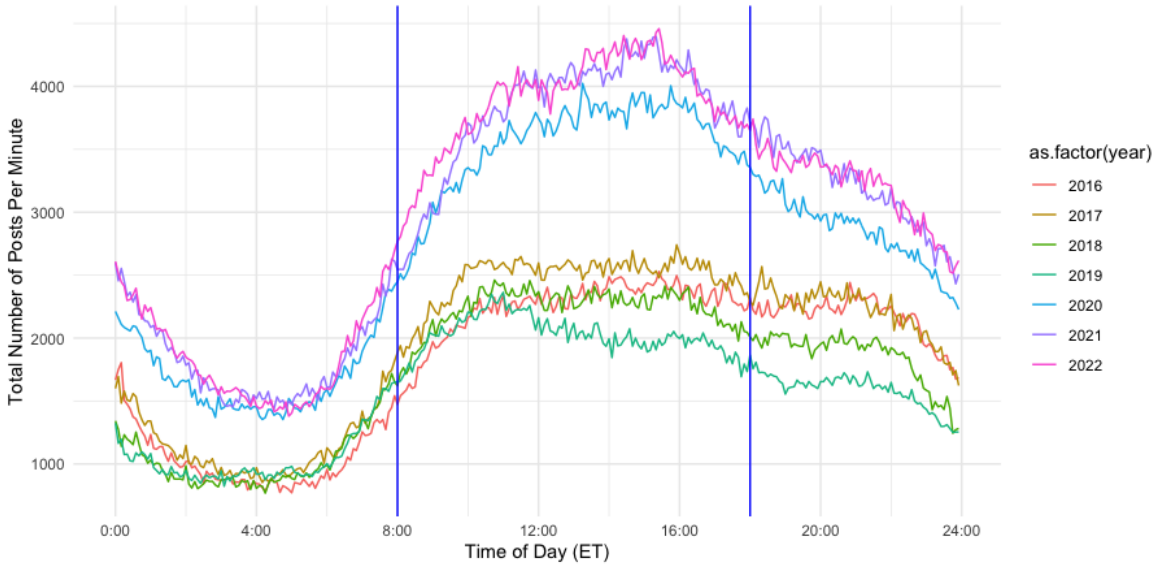


Figure 9: Distribution of posts across time of day (US Eastern Time). The figure shows the total number of EJMR posts per minute for a given year. The vertical blue lines show standard work hours for US Eastern time.

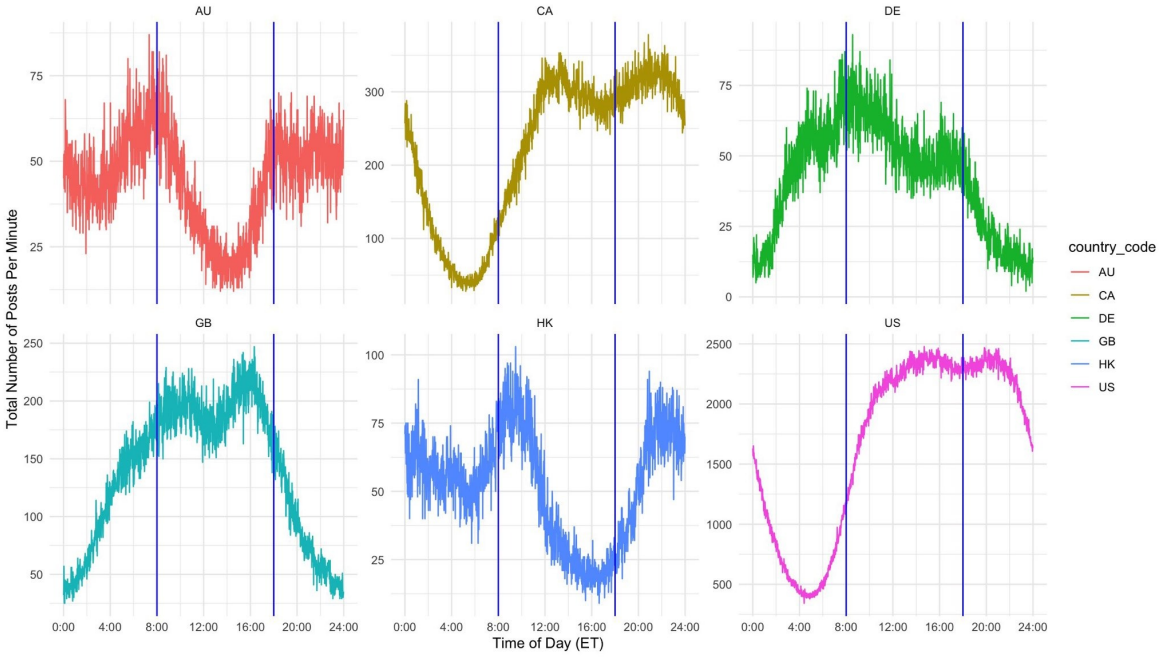


Figure 10: Distribution of posts across time of day (US Eastern Time) by country. The figure shows the total number of EJMR posts per minute for Australia, Canada, Germany, United Kingdom, Hong Kong, and the United States. The vertical blue lines show standard work hours for US Eastern time.

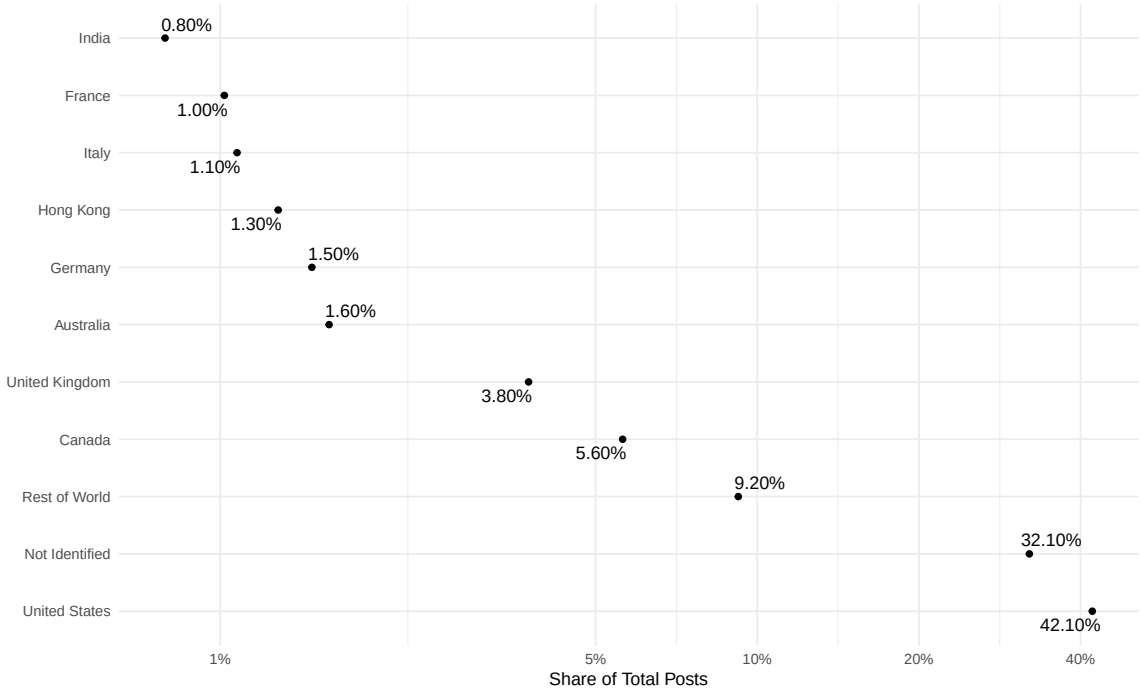


Figure 11: Distribution of posts across countries. The figure shows the share of all posts which can be assigned to a particular country. Posts for which we do not assign an IP address are in the Not Identified category.

	Language of Post					
Origin Country of Post	German	Chinese	Spanish	Portuguese	Russian	Korean
Germany	**0.80**	0.04	0.05	0.03	0.07	0.01
China	0.07	**0.85**	0.05	0.03	0.00	0.00
Hong Kong	0.09	**0.86**	0.03	0.02	0.00	0.00
Spain	0.29	0.03	**0.35**	0.33	0.01	0.00
Portugal	0.28	0.03	0.29	**0.40**	0.00	0.00
Brazil	0.09	0.00	0.12	**0.77**	0.01	0.00
Russian Federation	0.29	0.06	0.05	0.03	**0.57**	0.00
Korea	0.20	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 table 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.

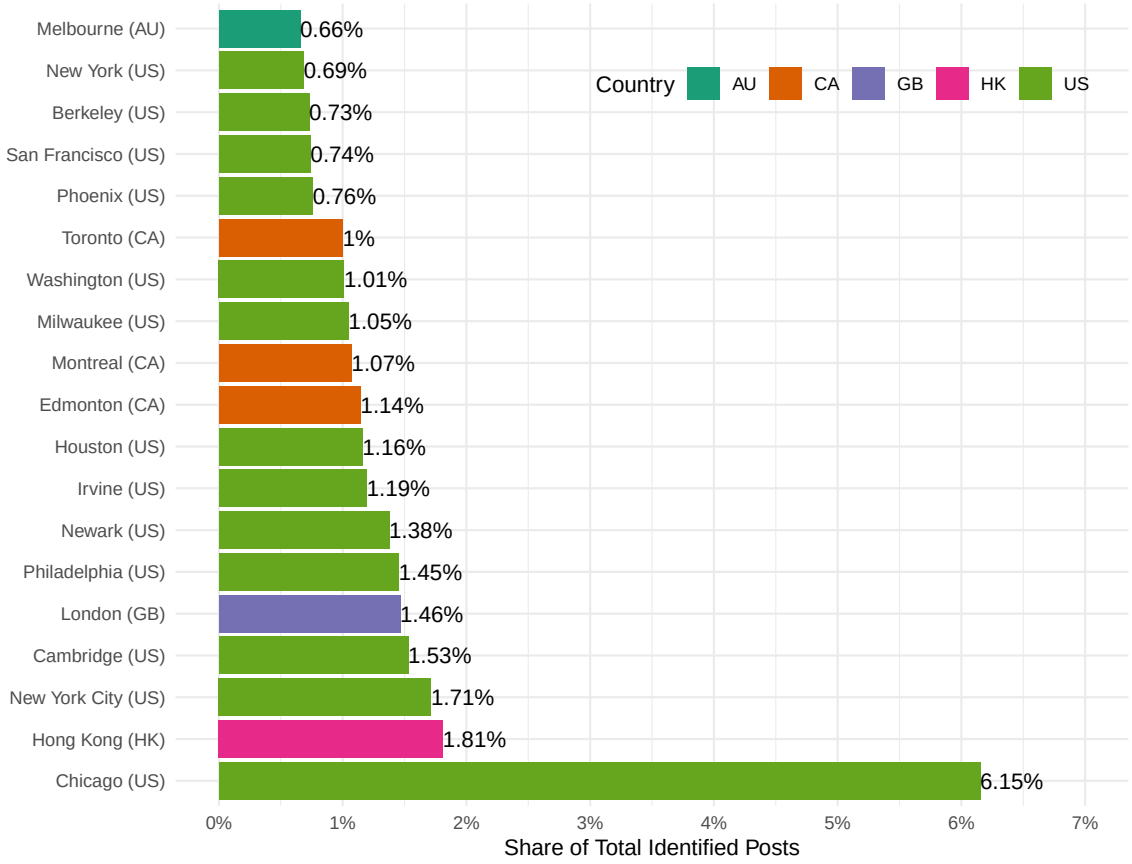


Figure 12: Share of posts with assigned IP address across cities. This figure shows the share of posts with an assigned IP address that originate from a given city. The share of posts from cities located in the United States, Hong Kong, the United Kingdom, and Canada are marked in pink, purple, orange, and green, respectively.

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 a 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 larger 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 top universities such as Cambridge or Berkeley are towards the top of the ranking despite their relatively small population size.

Of course, IP addresses reveal far more about their users than merely their geographical location. That said and as mentioned before, users change IP addresses. This can occur for several reasons such as dynamic IP address assignment or residential users restarting their internet routers. Most purchasers of business class internet have static IP addresses assigned to

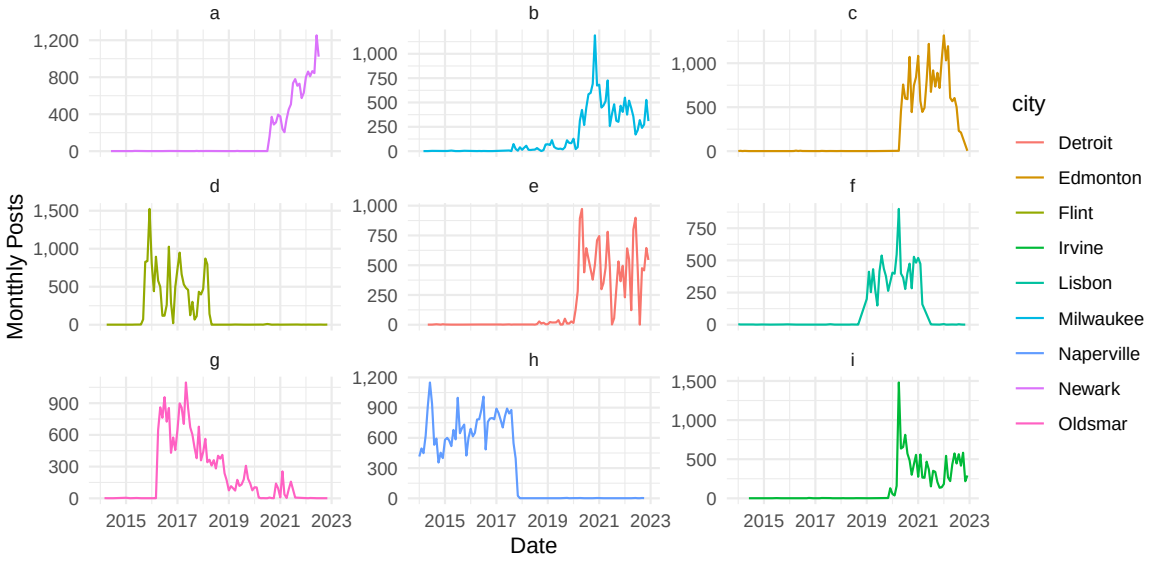


Figure 13: Monthly posts for selected EJMR power users. The figure shows the number of monthly posts over time for a select number of IP addresses from which a large number of posts originate.

them since servers and a number of business applications or Virtual Private Networks (VPNs) typically require the same IP address every day in order to operate properly. Figure 13 shows 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 start and 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).

Although posting on EJMR is generally frowned upon in the economics profession, 10.2% of all posts to which we assign IP addresses originate directly from IP addresses associated with universities or research 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 at their workplace. Perhaps even more surprisingly, there are EJMR posts from identified IP addresses located at *literally all* the leading universities in the United States. Posting on (and not just reading) EJMR thus appears to be pervasive and widespread, even from devices directly connected to university networks.

Figure 14 reports the share of each of the US universities with a top 25 economics department of all the posts originating from IP addresses associated with universities or research institutions. The figure particularly 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

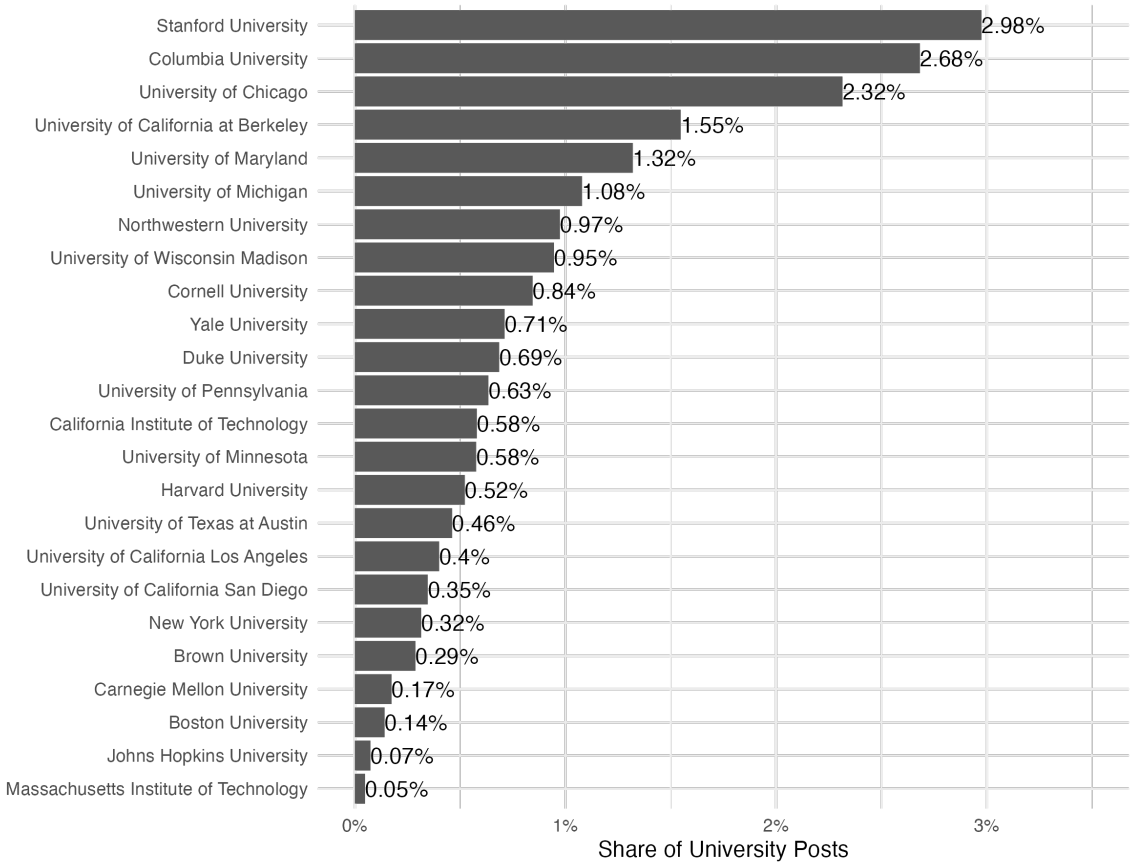


Figure 14: Post share of all university or research institution posts by each of US universities with a top 25 economics department. The figure shows the share of posts accounted for by a given top 25 US university among all posts originating from IP addresses associated with universities or research institutions.

than 20% despite there being several hundred other universities in our sample.

The sustantial overrepresentation of top US universities among EJMR posters is also apparent in Figure 15 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 and among the top four universities contributing to EJMR there are no less than three (Stanford, Columbia, and University of Chicago) which are ranked in the top 10.

However, the vast majority of EJMR posts originate from non-university IP addresses. Figure 16 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,

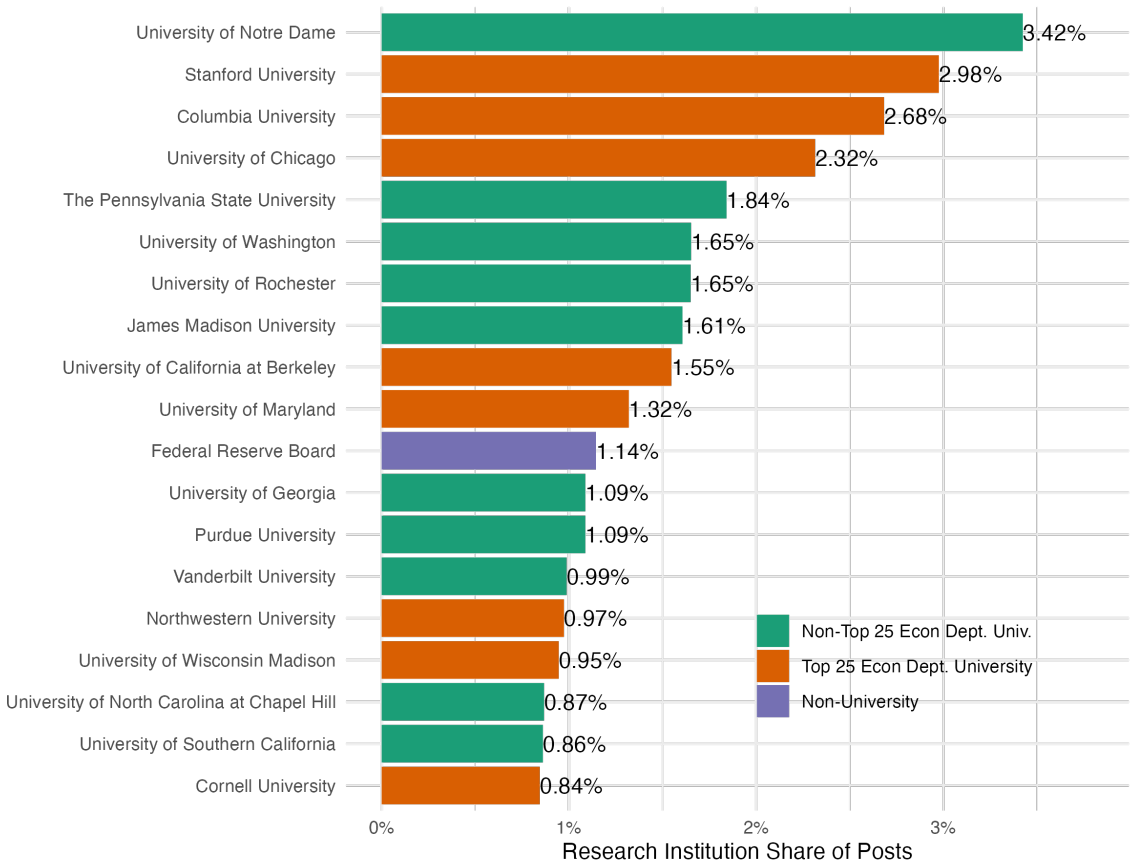


Figure 15: Post share of all university or research institution posts by largest universities and research institutions. The figure shows the share of posts accounted for by a US university or research institution among all posts originating from IP addresses associated with universities or research institutions.

providing suggestive evidence of a power law (Clauset et al., 2009). However, as we show in detail later, the curve slightly "bulges out" thus 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 three are served by university ISPs suggesting that power users are much more like to be posting from residential IP addresses. However, among IP addresses that post frequently but not as frequently as the top 100 IP addresses such as between rank 100 and rank 1,000, the majority of IP addresses is located at universities.

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

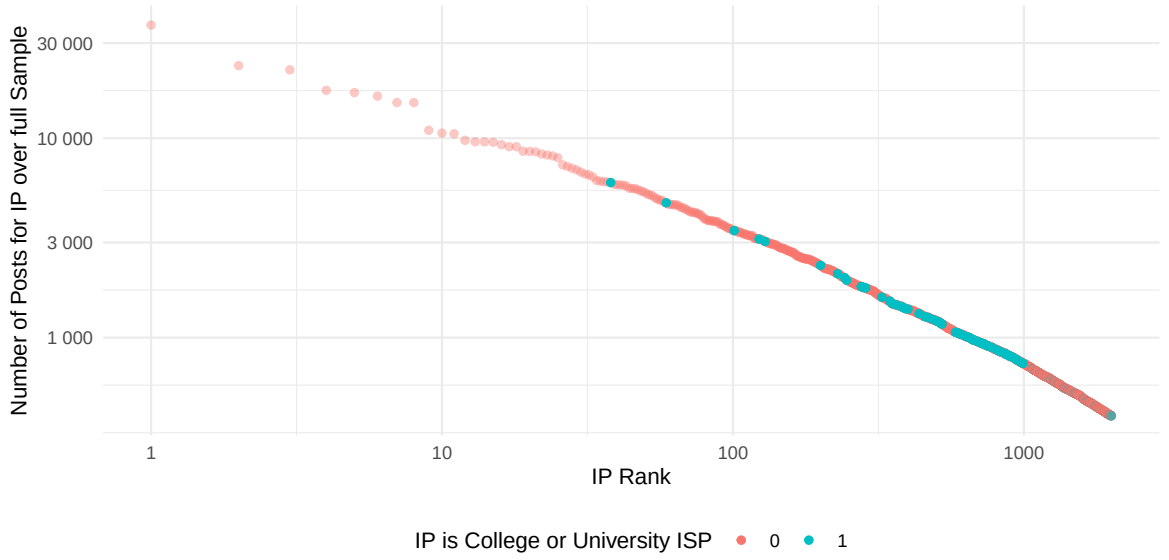


Figure 16: Number of posts by rank of IP address (according to number of posts) split by university and non-university ISP. Green dots indicate that the IP address is a university or research institution.

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.

Figure 17 reports the cumulative share of posts originating from a given proportion of IP addresses for all posts and for toxic posts with assigned IP addresses. The figure further underlines the extremely high concentration of posts which is also apparent in Figure 16. In particular, it shows that a mere 5% of the 47,630 IP addresses generate well over 50% of all posts with assigned IP addresses and 20% of IP addresses generate just over 80% of all posts.[16]

The degree of concentration is even higher than Figure 17 might suggest because it only considers posts with assigned IP addresses. Recall that for 33.9% of 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 shown in the figure.

There is considerable evidence 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 18 which plots the relationship between IP rank of posters and number of

[16]This type of concentration of contributions is quite common on online platforms (Guo et al., 2008, 2009).

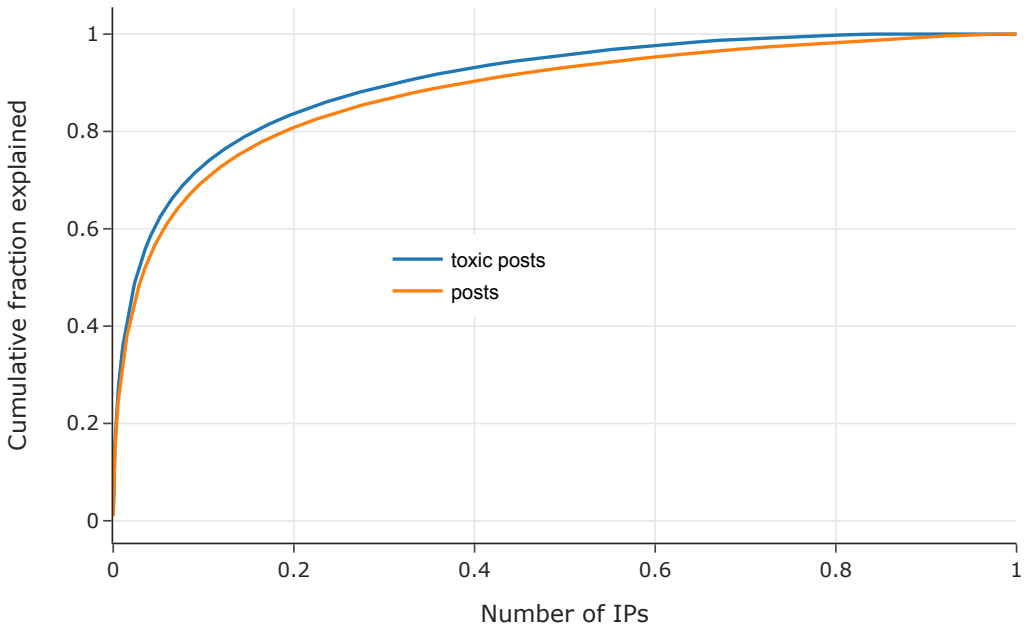


Figure 17: Cumulative share of posts originating from a given proportion of IP addresses for all posts and for toxic posts with assigned IP addresses. The figure shows what proportion of all posts (orange line) and toxic posts (blue line) with assigned IP addresses is generated by a given proportion of IP addresses. For example, 20% of IP addresses generate just over 80% of all posts with assigned IP addresses and almost 85% of all toxic posts.

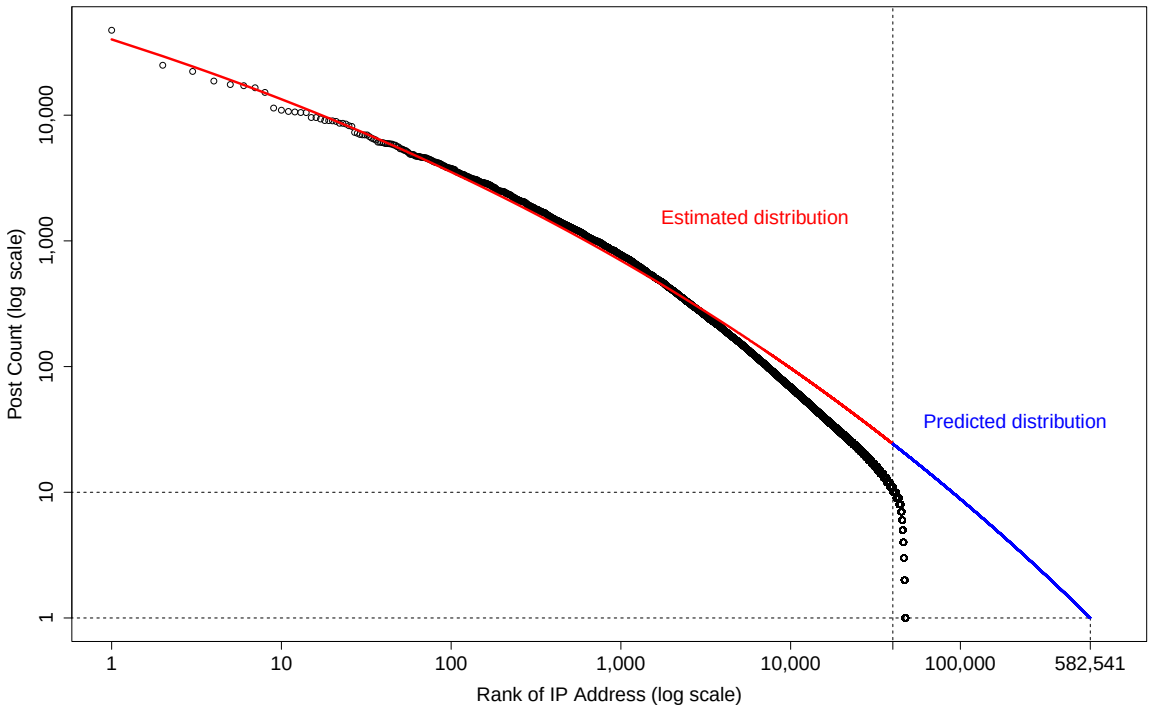


Figure 18: Distribution of posts by IP address rank. The figure shows the number of posts an IP addresses has contributed to EJMR. IP addresses are ordered on the x-axis by the number of posts assigned to them. We estimate a stretched exponential distribution (red line) of posts for the IP addresses 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.

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 10 topics in the span of a week it will not be assigned to any posts.

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 rank 40,000 which has 10 posts assigned to it and then projecting out this fitted distribution until implied number of posts of an IP address is equal to 1. The implied total number of posts is 7.4 million which matches the total number of observed posts which is equal to 7.1 million, quite well. Under this projection there are 582,541 IP addresses which have contributed at least one post to EJMR. While the vast majority of posts comes from just a few thousand IP addresses, 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, in particular how this content varies across universities and IP addresses.

3.5.1 Mentions of Universities

At its most useful EJMR is intended to be a source of information about the academic job market and professional news of the economics profession. Posts containing such perhaps useful information may be innocuous and and thus are more likely than toxic posts to come directly from within universities (i.e., from university IP addresses). 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 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. 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 Toxicity

As described in Section 2.5 we deobfuscate the content and then used a number of transformer-based machine learning models to classify posts by sentiment, misogyny, and toxicity. We also re-created all the word-count measures used in the seminal study of Wu (2020).

Figure 19 reports the share of toxic posts on EJMR over time as measured by the ToxiGen Roberta (Hartvigsen et al., 2022) which is a checkpoint of the Roberta model (Liu et al., 2019) fined-tuned for toxicity detection. The share of toxic posts has remained relatively constant at 10% over time. The graph exhibits some mild seasonality over the course of a year with a higher share of toxic posts during the summer months.

The concentration of posts across IP addresses is even higher for toxic posts than for all

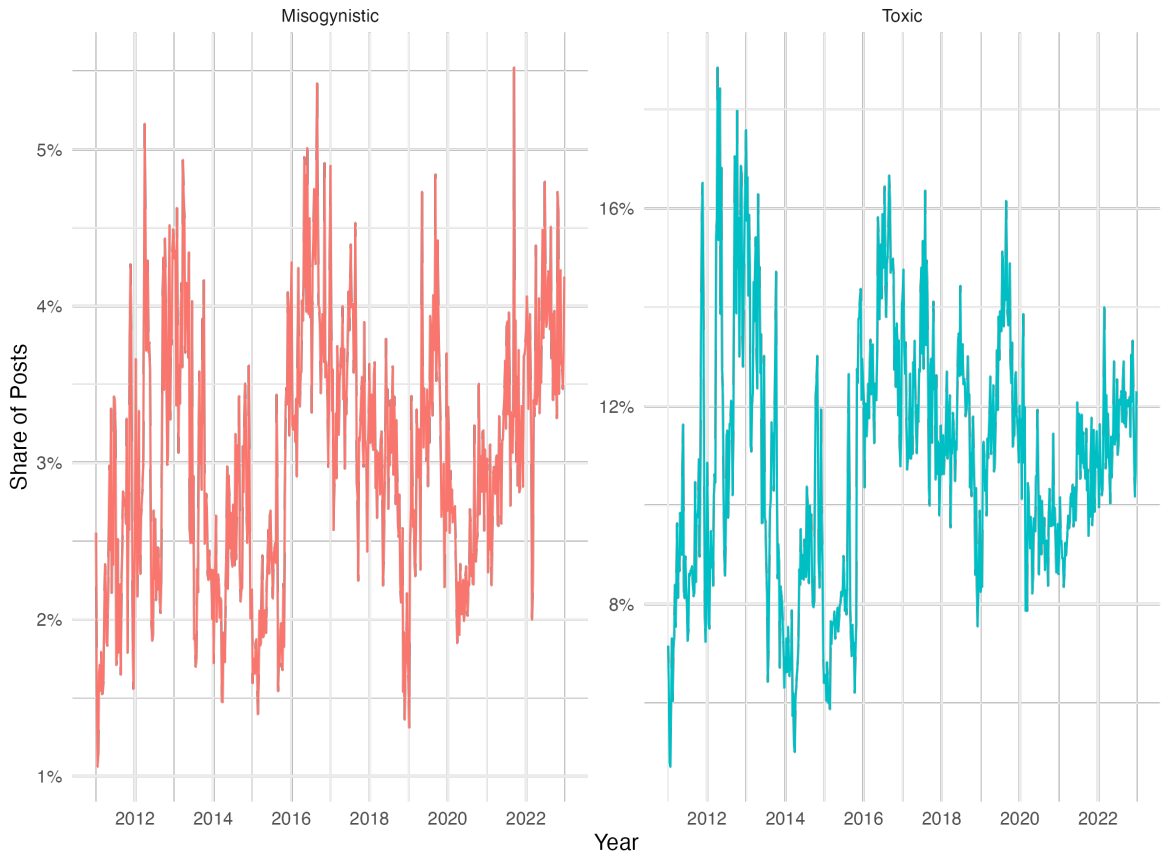


Figure 19: Share of toxic posts over time. The figure reports the share of toxic posts on EJMR over time as measured by the ToxiGen Roberta (Hartvigsen et al., 2022).

	Mentions of University									
University ISP	Harvard	MIT	Stanford	Berkeley	UChicago	Yale	NYU	NWU	Columbia	UPenn
Harvard	**7.9**	9.0	5.2	1.4	3.7	2.0	2.1	0.9	1.0	1.2
MIT	4.7	**9.8**	6.0	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	1.0
Yale	1.5	3.8	0.9	0.7	1.8	**3.4**	1.3	0.5	0.4	1.0
NYU	2.5	4.6	3.1	0.7	1.9	1.1	**5.8**	1.0	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	3.0	4.9	2.3	1.3	2.8	1.6	3.1	1.0	**5.0**	2.4
UPenn	2.2	3.5	1.8	0.9	2.4	1.0	3.1	1.1	0.3	**5.1**
Others	1.1	3.8	0.5	0.4	1.0	0.4	1.0	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"

posts. Figure 17 shows that less than 5% of IP addresses generate more than 60% of toxic posts.

The content of posts on EJMR varies across forums. Figure 20 shows the share of toxic posts by forum groups. Posts in the Off-Topic/Non-Econ forums are substantially more likely to be toxic than those in the Economics or Job Market Rumors forums. However, even in the Job Market Rumors forums in which discussion focuses on the academic job market, almost 7% of all posts are labeled toxic.

There is also considerable heterogeneity of the share of toxic posts across universities. Figure 21 lists the share of toxic posts for each of the universities among the top 25 economics departments in the United States according to the RePEc ranking.

One might expect that EJMR users exhibit greater inhibition when posting at work compared to when at home. If this were true, then we should observe that posts from university IP addresses are less likely to be toxic on average than posts from non-university IP addresses. Figure 22 shows that this is indeed the case. For all three groups of forums, posts coming from university IP addresses are less likely to be toxic than posts originating from non-university IP addresses.

In addition, 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 toxic university IP addresses there are several from leading US universities including the University of Rochester and the University of Chicago.

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 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.

Figure 24 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.

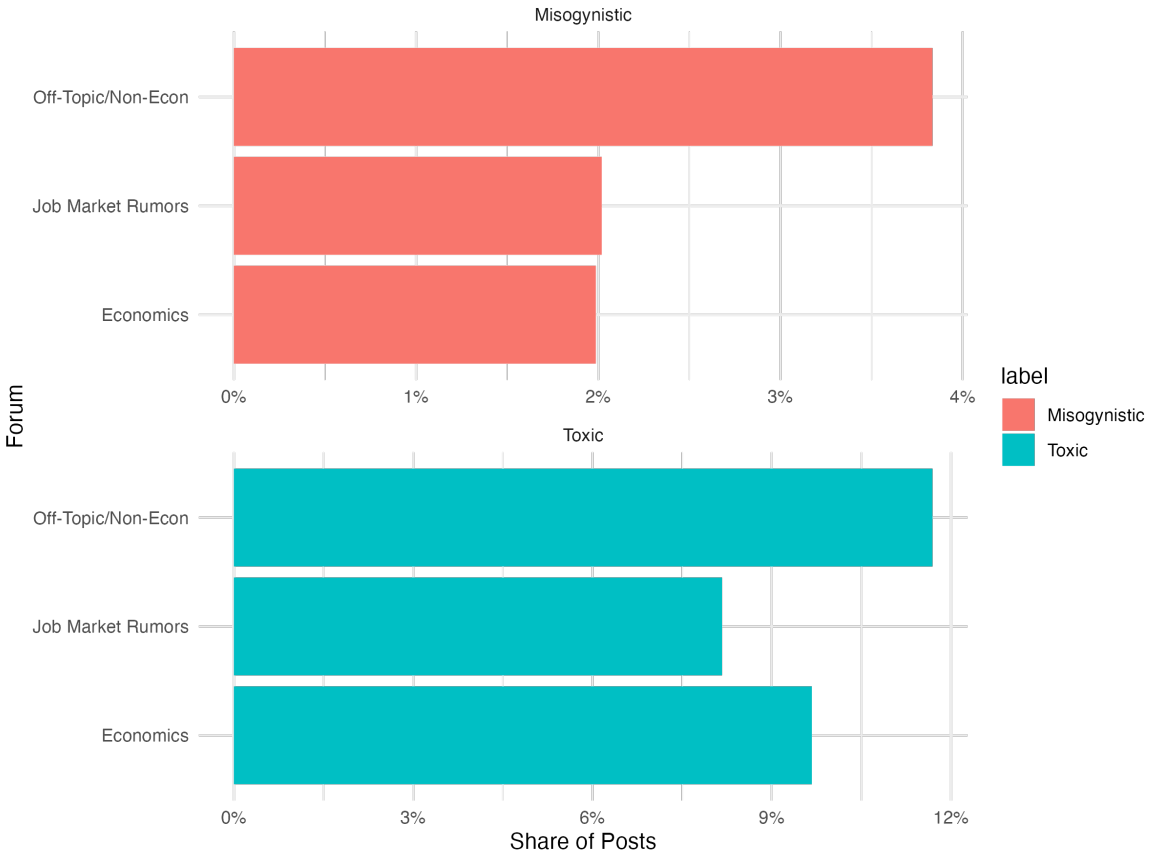


Figure 20: Share of toxic and misogynous posts by forum group. The figure reports the share of toxic and misogynous posts by forum groups as measured by the ToxiGen Roberta (Hartvigsen et al., 2022).

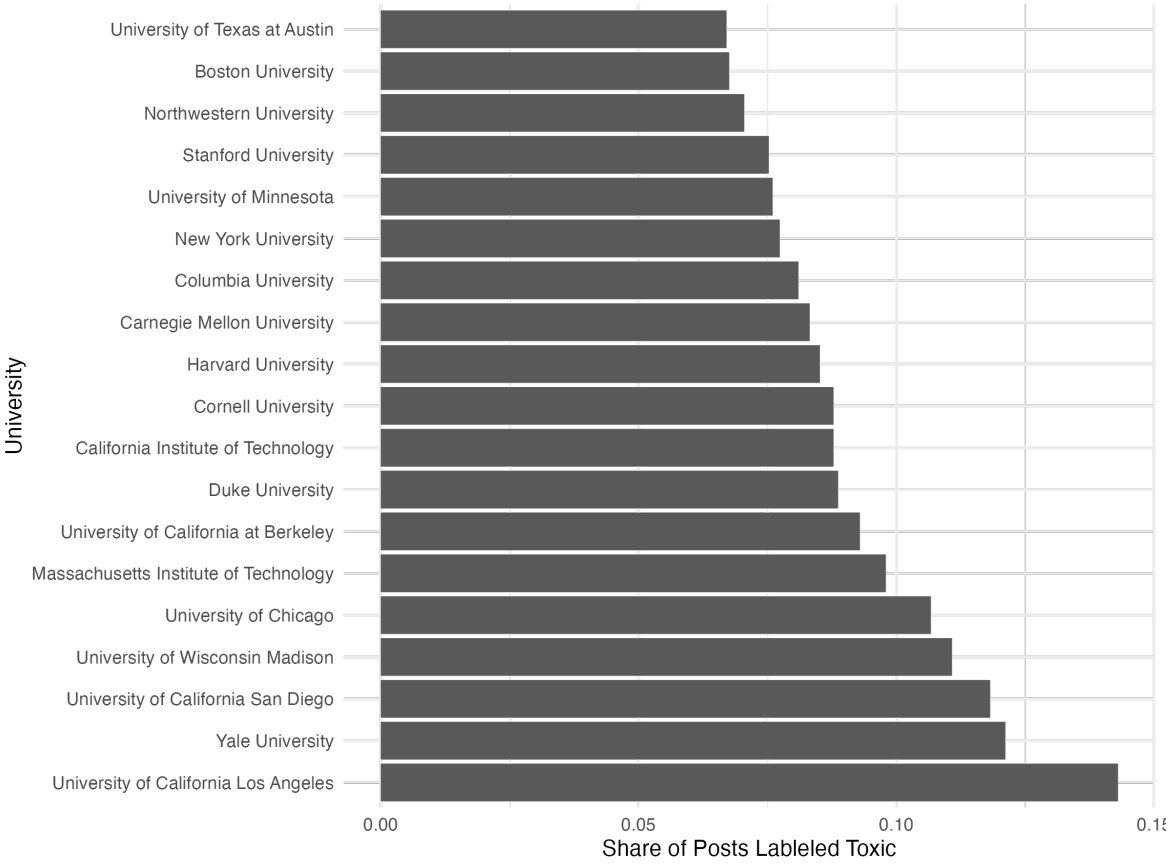


Figure 21: Share of toxic posts by university. The figure shows the share of toxic posts for each of the universities among the top 25 US economics departments.

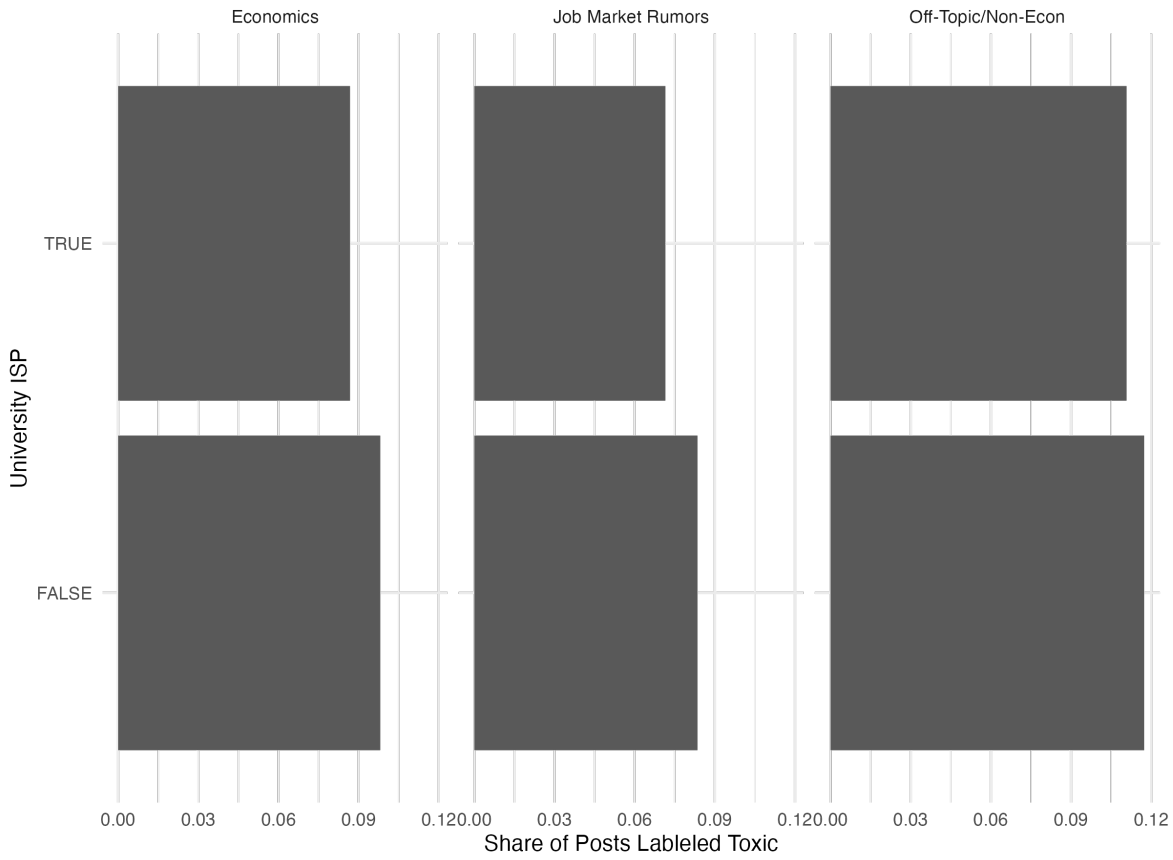


Figure 22: Share of toxic posts by university and forum group. The figure shows the share of toxic posts for each of the universities among the top 25 US economics departments and splits them by forum group.

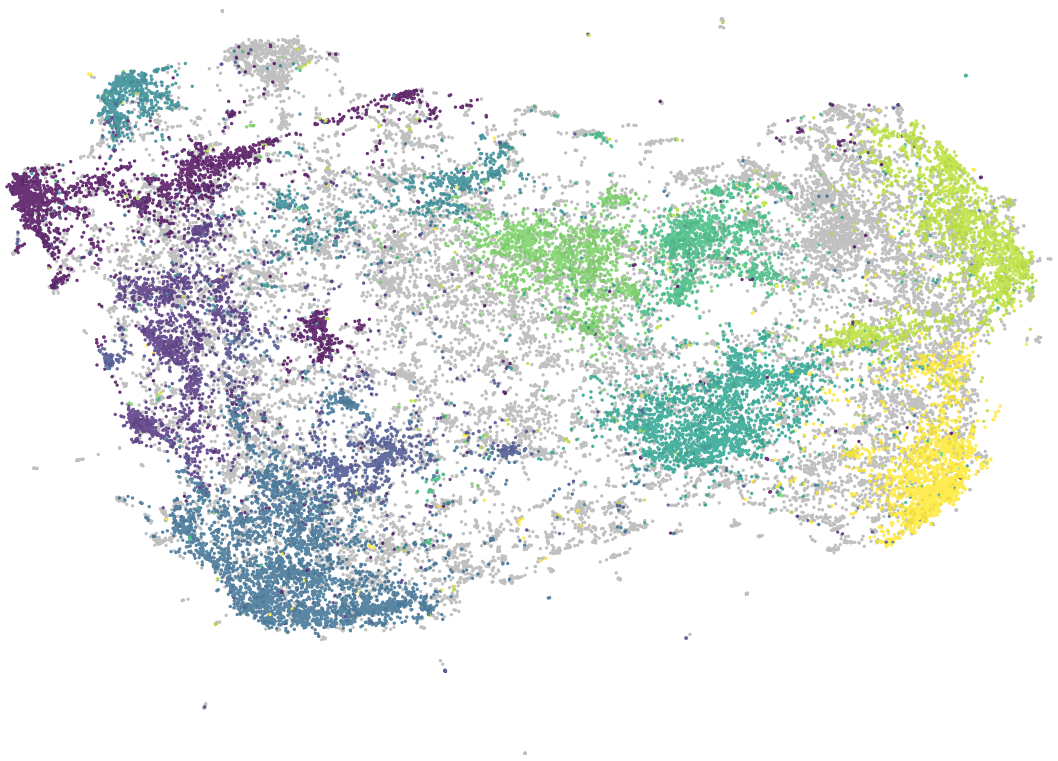


Figure 23: Linguistic patterns of EJMR contributing IPs. The chart shows one point for each IP address that contributed to EJMR in our data set (minus a few not meeting the criteria below). Proximity in the figure indicates linguistic similarity. To make this graph each post that was in English, at least 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.

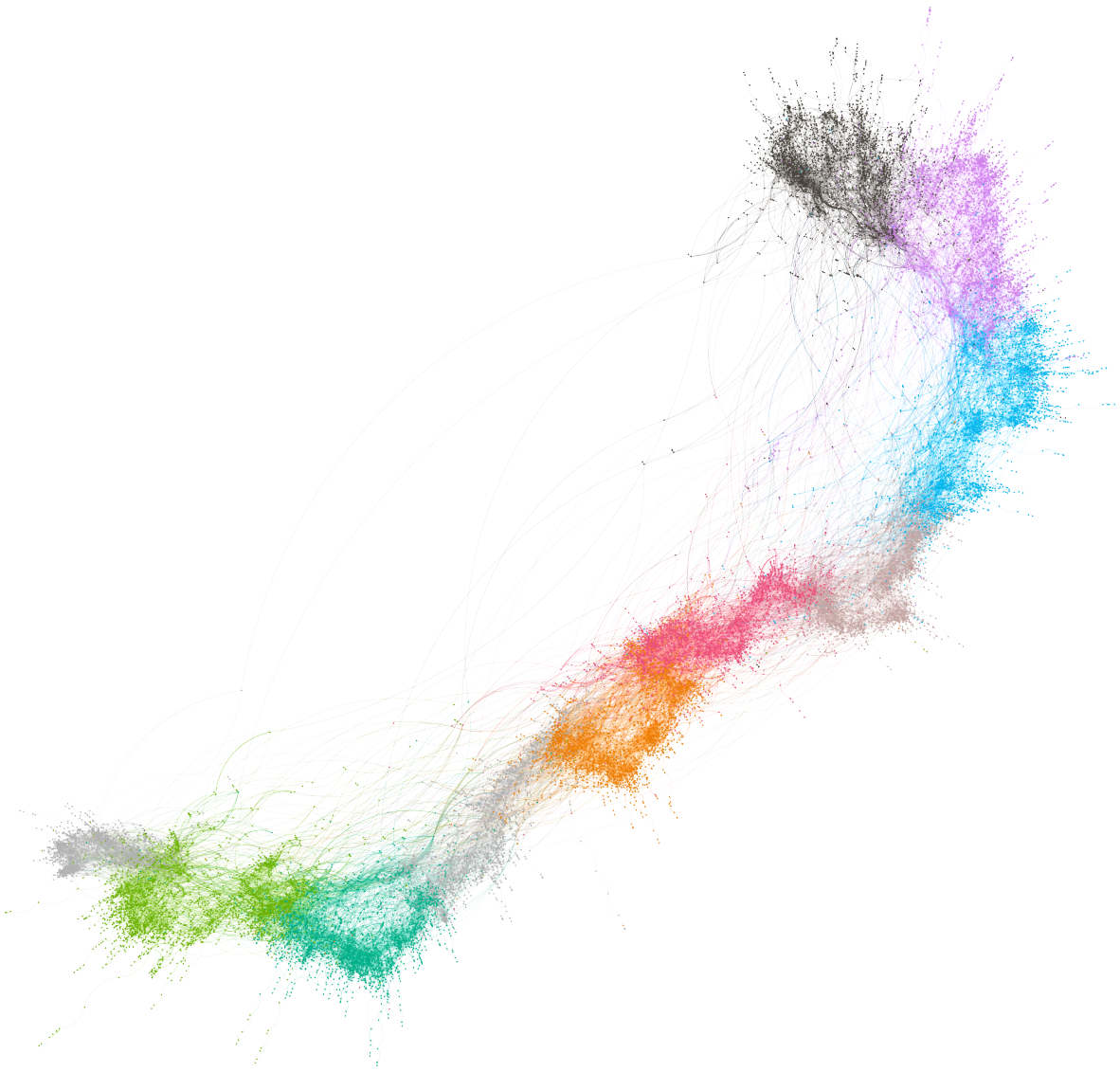


Figure 24: Interactions between IP addresses. The graph includes most of the IP addresses identified in our study with one vertex per IP address. Edges between vertices indicate that two IP addresses often posted in the same topics on EJMR. Color indicates mean year in which an IP address was active. 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.

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.

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.

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.

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*.

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.

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.