Journal of Complex Networks (2019) **0000**, 1–19
doi: 10.1093/comnet/cnz045

Modelling and analysis of tagging networks in Stack Exchange communities

XIANG FU[‡], SHANGDI YU[†,‡] AND AUSTIN R. BENSON
Department of Computer Science, Cornell University, 107 Hoy Rd, Ithaca, NY 14853, USA
[†]Corresponding author. Email: sy543@cornell.edu
[‡]These authors contributed equally to this work.

Edited by: Dr. Xiang Li

[Received on 27 July 2019; editorial decision on 13 November 2019; accepted on 16 November 2019]

Large Question-and-Answer (Q&A) platforms support diverse knowledge curation on the Web. While researchers have studied user behaviour on such platforms in a variety of contexts, there is relatively little insight into important by-products of user behaviour that also encode knowledge. Here, we analyse and model the macroscopic structure of tags applied by users to annotate and catalogue questions, using a collection of 168 Stack Exchange websites that span a diversity of sizes and topics. We study the distribution of tag frequencies and also the structure of 'co-tagging' networks where nodes are tags and links connect tags that have been applied to the same question. We find striking similarity in tagging structure across Stack Exchange communities, even though each community evolves independently (albeit under similar guidelines). Our findings thus provide evidence that social tagging behaviour is largely driven by the Stack Exchange platform itself and not by the individual Stack Exchange communities. We also develop a simple generative model that creates random bipartite graphs of tags and questions. Our model accounts for the tag frequency distribution but does not explicitly account for co-tagging correlations. Even under these constraints, we demonstrate empirically and theoretically that our model can reproduce a number of the statistical properties that characterize co-tagging networks.

Keywords: co-tagging networks; folksonomy; Stack Exchange.

1. Introduction

Question-and-Answer (Q&A) platforms are now a standard context for social interaction on the Web with platforms such as Quora and Stack Exchange supporting large user bases. As a result, the social networks that these platforms support have undergone a great deal of study, including, for example, how people find interesting and popular questions on Quora [1], prediction of 'best answer' selection on Yahoo Answers [2], market design for knowledge base construction with Google Answers [3] and badge collection on Stack Overflow [4]. These studies have largely focused on models and analysis of the user behaviour. However, the users also create richly structured data. In this article, we model and analyse the structure revealed by tags on Stack Exchange, which users employ to annotate and catalogue questions. Thus, our principal object of study is the tags (and their relationships through co-tagging), rather than the users.

A Stack Exchange website is a Q&A forum for a particular community. The platform began with Stack Overflow, which is a community for computer programming. Stack Overflow is the largest and arguably most well-known Stack Exchange community, but the Stack Exchange ecosystem supports a

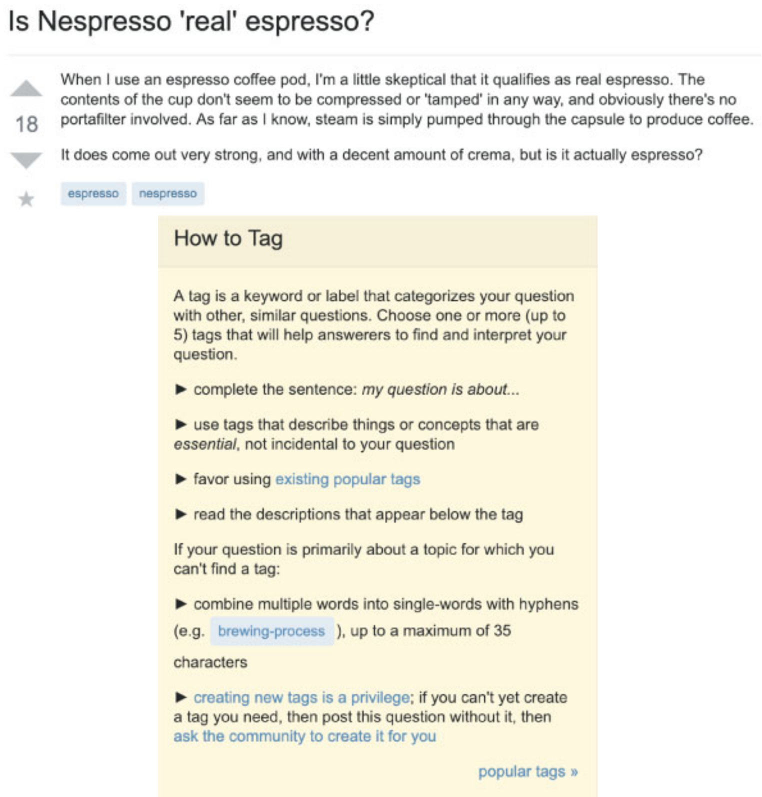

FIG. 1. Stack Exchange tagging. (*Top*) A question on the COFFEE Stack Exchange community with two tags: `espresso` and `nespresso` (from https://coffee.stackexchange.com/q/1572). We study the tag frequency distributions across a large collection of Stack Exchange communities, as well as networks constructed from tags applied to the same questions. (*Bottom*) User interface of tagging guidelines on the COFFEE Stack Exchange (https://coffee.stackexchange.com/questions/ask). The last rule says that users cannot immediately create new tags without due process; thus, tagging is fundamentally different from hashtags on other social media platforms such as Twitter or Instagram.

diverse set of communities ranging from pet ownership[1] to coffee[2] to philosophy.[3] For the most part, these Stack Exchange communities evolve independently under the same Q&A format (Fig. 1). A linchpin of every Stack Exchange community is the tagging system. When posting a question, users are encouraged to apply a small number of tags (at least one and at most five) that provide a reasonable abstraction of the question's topics. In addition to describing the question's content, tags also serve users in information retrieval of similar questions as well as questions they might be able to answer. Tags on Stack Exchange are not taken lightly—users cannot immediately create new tags and are encouraged to use existing and popular tags (Fig. 1, bottom); moreover, there are also official tagging guidelines.[4] Thus, tags on Stack

[1] https://pets.stackexchange.com/ (25 November 2019, date last accessed).
[2] https://coffee.stackexchange.com/ (25 November 2019, date last accessed).
[3] https://philosophy.stackexchange.com/ (25 November 2019, date last accessed).
[4] https://stackoverflow.com/help/tagging (25 November 2019, date last accessed).

Exchange are fundamentally different from, for example, hashtags on social media platforms such as Twitter which are largely free from regulation. The value placed on tags means that they can contain rich information about the community. For example, tag frequencies can show popular topics and the change of tag frequency over time can reveal the change of a community's interests over time. Here, we provide the first large-scale study of the macroscopic structure of tagging behaviour by analysing a collection of 168 Stack Exchange communities. We frame our study through the lens of network analysis, focusing on two networks constructed from the tagging behaviour of users. The first is the bipartite network of tags and questions, where there is an edge between a tag and all of the questions to which the tag was applied. The second is the co-tagging network, or the projection of the first network onto the tags; in this case, two tags are connected by an edge if the two tags jointly annotate at least one question. We also consider a weighted version of the second network, where the weight is the number of questions jointly annotated by the two tags.

Oftentimes, network analyses suffer from the fact that there is only 'one sample' of a social system to study. For example, there is only one Facebook friendship graph [5] and one Twitter follower network [6] to analyse. While such studies provide valuable insights into real-world social systems, it is also well known that there can be randomness in the evolution of social networks when crafted in a controlled setting [7]. Stack Exchange thus provides a unique opportunity to study a collection of similar networks of tags with highly similar dynamics that have evolved largely independently and differ most in the community topic (and implications of the community topic, such as the number of users).

We summarize our main contributions and findings as follows:

(1) *Empirical analysis on tag frequencies.* Across our collection of 168 Stack Exchange communities, we find that the distribution of tag frequencies is heavy tailed and well approximated by a lognormal distribution. The two parameters of this distribution are themselves well approximated by a normal distribution when estimated over the large collection of Stack Exchange communities. This finding provides evidence that basic tagging behaviour—at a high level—is independent of the Stack Exchange community and rather a property of the general Stack Exchange platform.

(2) *Empirical analysis on co-tagging networks.* The 'co-tagging network' is induced by the bipartite tag-question network. Specifically, we analyse the graph where the nodes are tags and there is an edge connecting two tags if they are 'co-tagged' on at least one question (with possible weighting on edges corresponding to the number of questions on which the two tags appear).

Our analysis focuses on three macroscopic properties of the networks. First, the weighted number of co-tags of a given tag is well approximated by a linear function of the number of questions in which the tag appears. Second, the number of unique co-tags of a given tag is well approximated by a simple third-degree polynomial of the number of the tag frequency. Qualitatively, as we increase the number of questions that a tag has appeared in, the number of unique co-tags will also increases; however, this growth tapers for popular tags, when it is difficult to accumulate more unique co-tags. Again, the similarity in structure across Stack Exchanges provides evidence that social tagging behaviour is more a function of the Stack Exchange platform itself and not of the individual Stack Exchange communities. Third, we measure three versions of the clustering coefficient for weighted and unweighted networks networks and find various levels of clustering and find that the unweighted clustering coefficient is only mildly correlated with the size of the Stack Exchange community (as measured by the number of questions), but two versions of the weighted versions both negatively correlate with size.

(3) *A random graph generative model.* Using our empirical findings as a guide, we devise a simple generative model for creating random bipartite graphs with links connecting tags to questions. The model takes as input the desired number of questions, number of tags, number of total tag occurrences, and two parameters of a lognormal distribution, and produces as output a bipartite graph linking tags to

questions. Despite the simplicity of this model, all three macroscopic properties of the co-tagging network are replicated by our model, which we validate with both empirical and theoretical analysis across the collection of 168 Stack Exchange networks. Importantly, the model does not bake in any notions of correlation or clustering in the co-tagging but can still replicate important co-tagging network properties. Thus, we can conclude that these network properties could actually be explained by our simple generative model that only makes a strong assumption on the frequency distribution of the tags. These findings contrast sharply with traditional social network analysis in measuring clustering. Standard random graph models for social networks that do not bake in clustering structure do not exhibit the same clustering levels as the real-world social system [8]. However, in our case, the co-tagging network constructed from our bipartite tag-question generative model matches the clustering levels in the empirical data. This is the first generative model for Q&A platforms that recovers the properties of the co-tagging network.

2. Related work

We now summarize related research in social media, information retrieval and network science.

2.1 *Online Q&A platforms and social media tagging*

Question-and-Answer (Q&A) platforms have been a staple of online discussion for several years, involving major web companies such as Yahoo!, Google and Quora. Research on these platforms has spanned a variety of topics, including reputation mechanisms [9, 10], answer quality measurement [1, 11, 12], network structure [2, 13]; social behaviour [14]; answer prediction [2, 15]; topic popularity [16]; and expertise evaluation [12, 17, 18]. This research has largely focused on the questions, answers and user behaviour. Our article, in contrast, treats tags as the fundamental object of study. Furthermore, most prior work has only examined at most a few Q&A web sites, whereas we study a large collection of Stack Exchange networks.

There are core differences in tagging on Stack Exchange compared to other web sites (as mentioned in Section 1), and this can have an effect on the tag semantics and user behaviour. One key difference is that Stack Exchange has a limit of five tags, so users are often forced to focus on the main topics of the question and not the minor details. In contrast, on Quora, around 20% of questions are annotated with more than five topics [1]. Another key difference is in tag creation—Stack Exchange encourages tag re-use but other social media platforms (e.g. Twitter) permit free tagging. This leads to users applying tags on Stack Exchange primarily to catalogue question content, rather than as a mechanism for augmenting content, as is common on Twitter [19]. This can be seen in tag re-use, which is often low on Twitter [20].

2.2 *Folksonomy*

The tag-question network that we study is related to the idea of *folksonomy*, a term coined by Thomas Vander Wal to describe the practice of users tagging information for personal retrieval in an open social environment [21]. Folksonomy has been a lens for analysis on social media platforms such as `CiteULike`, `del.icio.us`, and `BibSonomy` [22–24]. A major difference of these folksonomy studies and the present work is that folksonomies are much less restricted in the annotations—users can add many (possibly new) annotations freely—whereas the Stack Exchange system is restricted (between one and five tags with systematic vetting of new tags). And again, we analyse a large collection of Stack Exchange communities and not just a few folksonomies.

2.3 *Bipartite network models and co-tagging networks*

Bipartite graph (network) models are used in a broad range of scientific disciplines, including ecology [25], biomedicine [26] and information science [27]. The model that we develop in this article is a *generative (random) model* for a bipartite graph (network) between tags and questions. Other generative models for bipartite (or multipartite) graphs include the bipartite stochastic block model [28], evolutionary affiliation networks [29] and generative models for folksonomy [30]. In contrast to prior research, the goal with our model is to develop a simple generative model that captures the empirical properties that we observe to persist across Stack Exchange communities. Our model is designed to capture the tag frequency distribution amongst questions, but we find that properties of the co-tagging network—where tags are connected if they have appeared in a question together—are still replicated with our model. Properties and statistics of co-tagging networks, such as clustering coefficients, characteristic path lengths and number of co-tags have been used to analyse online communities such as `del.icio.us` and `BibSonomy` [24, 31]. Co-tagging networks have also been used for application on connecting users with similar interests [32].

3. Data description and preliminary analysis

A Stack Exchange is a self-moderating online Q&A forum, and each Stack Exchange community centres on a different topic. Questions are annotated with at least one and at most five tags that serve as essential descriptors of the question (Fig. 1). Importantly, these platforms also largely evolve independently, allowing us to perform a better statistical analysis compared to analysing a single Stack Exchange community. We now describe our dataset collection and provide preliminary statistical analyses that will serve the development of our generative model later in the article.

3.1 *Data collection*

We collected data from `https://archive.org/details/stackexchange`, which hosts the entire history of every Stack Exchange community, including the tags used to annotate questions. In total, we collected the sets of tags applied to each question from 168 Stack Exchange communities. In order to ensure that we could analyse data by inspection, we omitted communities where the predominant language was not English (thus, we do not consider the ES, JA, PT, RU, RUS and UKRAINIAN communities). However, we do include Stack Exchange communities such as RUSSIAN, where people discuss the Russian language in English. We also omitted 'meta' communities that discuss a particular Stack Exchange community since these meta communities have a different set of goals as well as a dependence on the community that they discuss. Finally, we also omitted Stack Overflow, which is over an order of magnitude larger than any other community, and has already been the subject of much research [1, 9, 11].

More formally, each Stack Exchange community consists of a set of posts or questions Q, and each question $q \in Q$ has an associated set of tags T_q. The structure of Stack Exchange enforces that $1 \leq |T_q| \leq 5$. We will often consider the set of *unique tags* for a community, that is, $\cup_{q \in Q} T_q$ (since Stack Exchange does not allow the free creation of tags, the set of unique tags can be thought of as a more static set of descriptors that can be applied to questions).

Figure 2 presents an overview of the basic statistics of our collection of tags. Among the 168 Stack Exchange communities in our analysis, the number of unique tags ranges from 70 (ARABIC) to 5,318 (SUPERUSER), and the number of questions ranges from 122 (again, ARABIC) to 994,983 (MATH). Although the Stack Exchange communities vary in size and topic and also evolve largely independently, we see in the next section (and later in the article) that there are broad similarities across the communities.

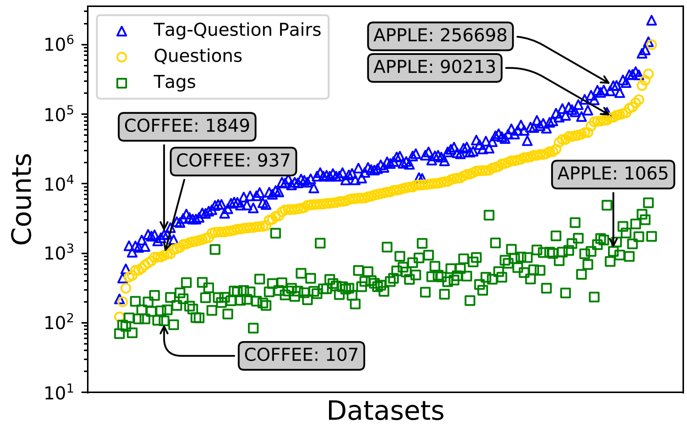

FIG. 2. Number of unique tags (green squares), questions (yellow circles) and tag-question pairs (blue triangles) of the 168 Stack Exchange communities analysed in this article. Datasets are sorted in ascending order by number of questions. The APPLE and COFFEE communities are annotated as examples. In this article, we analyse and model the relationships between tags and questions.

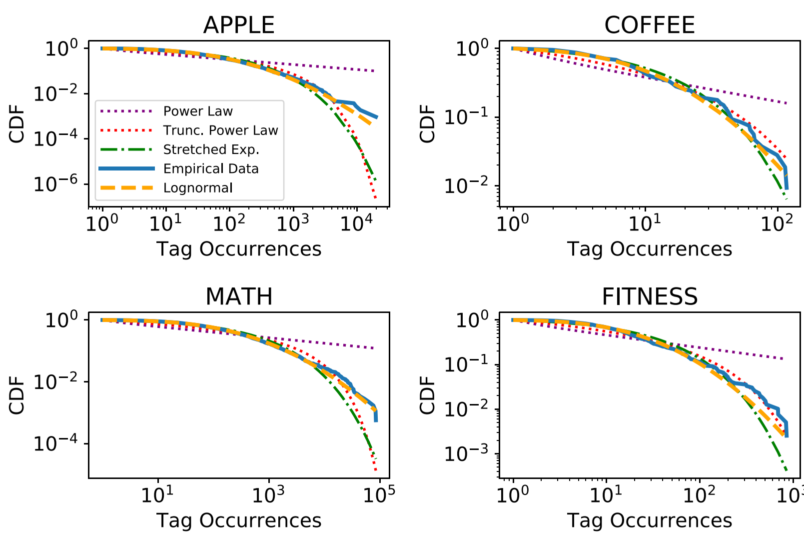

FIG. 3. Tag frequencies of four diverse Stack Exchange communities (APPLE—90,213 questions and 1,065 tags; COFFEE—937 questions and 107 tags; MATH—994,983 questions and 1,749 tags; FITNESS—7,626 questions and 393 tags). We find that tag frequencies are well modelled by a lognormal distribution in Stack Exchange communities (see also Fig. 4).

3.2 *Lognormal distribution of tag frequencies*

In this section, we study the distribution of tag frequencies, that is, the number of times that a tag is applied to a question or, when normalized, the fraction of questions that contains a given tag. One consistent trait is that tags used only a few times are much more common than tags used many times, and the distribution of tags is heavy tailed. Many communities have tags appearing at much higher frequencies than most other

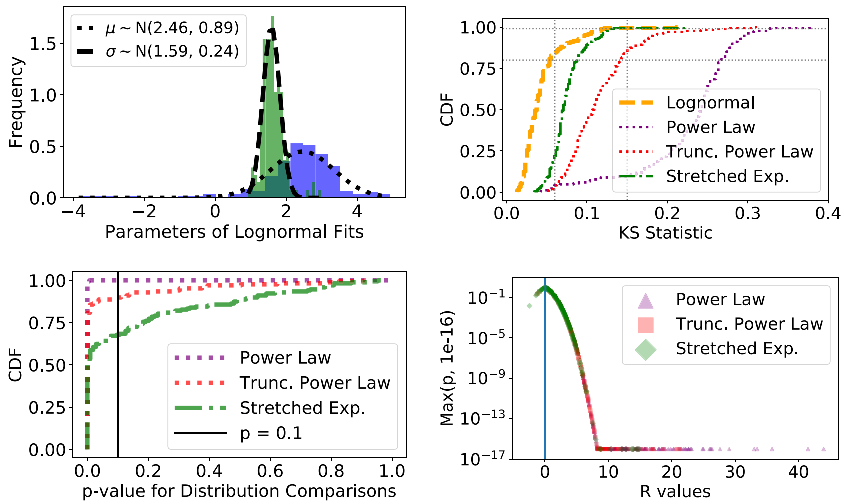

FIG. 4. (*Top left*) The distribution of fitted lognormal parameters for tag frequency across 167 Stack Exchange communities (we exclude the outlier PATENT), which are approximately normally distributed. (*Top right*) The CDF of the KS Statistic (D) of fitted heavy-tailed distribution to the empirical data. The lognormal distribution has the smallest statistics, and $D < 0.06$ for 80% of datasets; the only community with $D > 0.15$ is PATENT. (*Bottom*) The p-values for comparing an alternative heavy-tailed degree distribution hypothesis to the null hypothesis of the lognormal (left) and the distribution of the p-value with the log-likelihood ratio R (right). The small p-values and positive log-likelihood ratios indicate that the lognormal is a better fit for the data compared to other common heavy-tailed distributions.

tags; as an extreme example, the `magic-the-gathering` tag appears in more than 3,000 questions in the BOARDGAMES community, while all other tags appear in fewer than 500 questions.

Such heavy-tailed distributions are common on the Web and other domains [33, 34]. Here, we find that the tag frequencies are well modelled by a lognormal distribution. Figure 3 illustrates four representative cases and also provides a comparison against other commonly used heavy-tailed probability distributions such as a power law, truncated power law and stretched exponential. We find that a lognormal tends to match both the head and tail of the distribution, while other common heavy-tailed distributions can only capture either the head or tail of the distribution (e.g. in Fig. 3, the truncated power law captures the head of the tag frequency distribution in APPLE but not the tail and the tail of the COFFEE distribution but not the head). The lone outlier is the PATENT community, which does not seem to be well approximated by any commonly used heavy-tailed distribution.

More formally, we fit the parameters of a lognormal, power law, truncated power law and stretched exponential distributions to the tag frequency of each Stack Exchange community using the `powerlaw` Python package [35]. Figure 4 (top left) shows the fitted parameters, which are themselves approximately normally distributed. We use two standard procedures for evaluating the fit of the lognormal: the Kolmogorov–Smirnov (KS) statistic and the likelihood ratio test comparing the lognormal to other heavy-tailed degree distributions [33]. The distribution of the KS statistics is much smaller for the lognormal compared to the other distributions (Fig. 4, top right) and is less than 0.06 for 80% of the Stack Exchange communities. Furthermore, the p-values from the likelihood ratio test show that the power law, truncated power law and stretched exponentials are not likely alternatives to the null of a lognormal (Fig. 4, bottom). We performed a similar analysis for the distribution of the number of *co-tags* over tags (two tags are co-tags if they appear in some question together). We found a similar result in that these other heavy-tailed distributions are not likely alternatives to the lognormal; however, we also found that the lognormal was not a great fit for the data itself. Thus, we focus on the more universal characteristics of the tag frequency distributions.

To summarize, a lognormal distribution is an appropriate model for the distribution of tag frequencies. The fact that this distribution is appropriate across 167 of 168 stack exchange communities indicates that the tag frequency distribution is a property of the Stack Exchange platform at large and not of the individual communities themselves. Following this finding, we describe in the next section a simple generative model for random bipartite graphs of tags and questions based on this lognormal distribution. We will then later see that this model matches the real data in a number of characteristics related to co-tagging, that is, how multiple tags are used on the same question.

4. A generative model for bipartite tag-question networks

In this section, we propose a simple generative model for the bipartite tag-question network. Later, we will see that this model is able to recover many properties of the co-tagging network of Stack Exchange communities, that is, the graph where nodes correspond to tags, and edges connect tags that have been applied to the same question. Formally, the bipartite tag-question graph B consists of disjoint vertex sets T and Q, each corresponding to the set of tags and questions, as well as a set of undirected edges E; where $(t, q) \in E$ with $t \in T$ and $q \in Q$ signifies that tag t is applied to question q. The frequency, or number of occurrences, of a tag t is then simply the degree of t in the graph B.

Our random network model has two basic steps. First, given $N_T = |T|$, $N_Q = |Q|$, and the parameters μ and σ of a lognormal distribution, we first generate a sequence of tag occurrence counts $x_t \sim \text{Lognormal}(\mu, \sigma^2)$. These samples are scaled by a constant so that $\sum_t x_t = m$ (where m is the total number of tag occurrences in the original dataset) and then rounded to an integer. Since scaling a lognormal random variable by a constant is still lognormally distributed, we maintain this property of the tag distribution, and this also preserves the total number of tag-question pairs in the dataset. Second, we assign tag t to x_t questions chosen uniformly at random without replacement. In this simplified version of the model, the output deviates from the Stack Exchange networks in two ways: (i) it is possible that a question has no tags and (ii) it is possible that a question is assigned more than five tags. We now show how to account for these deviations, and Algorithm 1 describes the full procedure.

4.1 *Correction for question counts*

To fix the problem where questions can have no tags, we make a ‘correction’ in the number of questions. More specifically, we increase the number of questions from N_Q to $\hat{N}_Q$ so that after the random assignment, the expected number of questions with at least one tag is close in expectation to N_Q, the number of questions in the empirical dataset. We then simply discard questions with no tags (Algorithm 1).

We approximate the expected number of questions with no tags under a simplification where tags can be duplicated in questions (the approximation is not necessary, but it makes the calculations simpler, has small variance theoretically, and provides good results empirically). Here, the probability that a question gets 0 tags is the same for each question—it is just the probability that all tags are assigned to the other $\hat{N}_Q - 1$ questions. Formally, this is

$$\prod_{i=1}^{N_T} \prod_{j=0}^{x_i - 1} \left[1 - 1/(\hat{N}_Q - j)\right] \approx (1 - 1/\hat{N}_Q)^m, \tag{4.1}$$

Algorithm 1: Simple generative model for random bipartite graphs of tags and questions.

Input: number of tags N_T; number of questions N_Q;
target number of tag occurrences m; μ, σ^2
Output: tag-question bipartite graph $B = (T \cup Q, E)$
`/* Sample tag occurrences and compute corrections. */`
$x'_t \sim \text{LogNormal}(\mu, \sigma^2)$, $t = 1, \ldots, N_T$.
$x_t \leftarrow \text{round}(m \cdot x'_t / \sum_{t=1}^{N_T} x'_t)$, $t = 1, \ldots, N_T$.
Solve $\hat{N}_Q - \hat{N}_Q \exp(-m/\hat{N}_Q) = N_Q$ for $\hat{N}_Q$.
$\hat{N}_Q \leftarrow \text{round}(\hat{N}_Q)$.
`/* Construct bipartite graph */`
$T \leftarrow \{1, \ldots, N_T\}$, $Q \leftarrow \{1, \ldots, \hat{N}_Q\}$.
for *each tag* $t \in T$ **do**
  $Q_t \leftarrow$ uniform sample of x_t questions from Q.
  for $q \in Q_t$ **do** add edge (t, q) to edge set E.
end
$Q \leftarrow \{q \in Q \mid \exists t \in T \text{ for which } (t, q) \in E\}$

where m is total number of tag occurrences. Thus, since $\hat{N}_Q$ and m are generally large, when assigning tags uniformly at random to $\hat{N}_Q$ questions, the expected number of questions with 0 tags is

$$\hat{N}_Q(1 - 1/\hat{N}_Q)^m \approx \hat{N}_Q \exp(-m/\hat{N}_Q). \tag{4.2}$$

There are N_Q questions if the following equation is satisfied:

$$\hat{N}_Q - \hat{N}_Q \exp(-m/\hat{N}_Q) = N_Q. \tag{4.3}$$

We claim that Eq. (4.3) has a unique positive solution $\hat{N}_Q > N_Q$. Since m and N_Q are positive constants, the left-hand side of Eq. (4.3) is a function f of $\hat{N}_Q$. Moreover, the function f is continuous and monotonically increasing in $\hat{N}_Q$, and $f(N_Q) = N_Q(1 - \exp(-m/N_Q)) < N_Q$. Therefore, the above equation has a unique positive solution for $\hat{N}_Q$ that is larger than N_Q. We can find the solution efficiently with binary search, and then round $\hat{N}_Q$ to the nearest integer.

In our experiments, using the corrected number of questions with our model is accurate, even with our approximations. Generating one sample for each dataset, the relative error between the number of questions with at least one tag in the model deviates from the true number of questions by 0.32% on average and by at most 3.75% across all datasets. While these statistics are for just one sample in each network, the variance in the number of questions with 0 tags is approximately $\hat{N}_Q p(1 - p)$. The ratio between the theoretical standard deviation and the corrected number of questions is small—less than 0.008 for 80% of the datasets (Fig. 5, left).

4.2 *Number of tags per question*

We next justify our second model deviation, which is that questions can be assigned more than five tags. Our argument is simply that only a small fraction of questions are actually assigned more than five

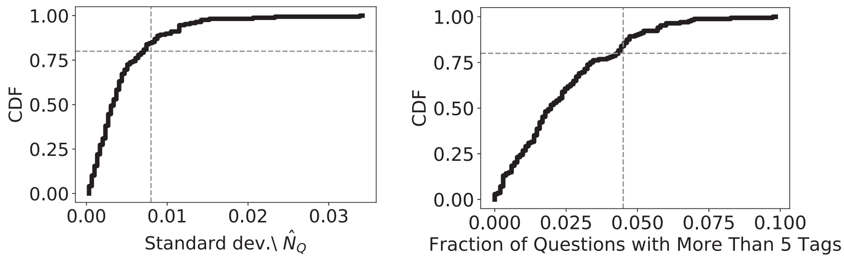

FIG. 5. (*Left*) The CDF of the ratio of the (theoretical) standard deviations to the corrected number of questions. The small ratio (less than 0.008 for 80% of datasets) shows that our correction for the number of questions is accurate. (*Right*) The CDF of the fraction of questions with more than five tags in one sample of the random graph for each dataset. This fraction is small—less than 0.045 for 80% of datasets, which justifies the relaxation in our random graph model.

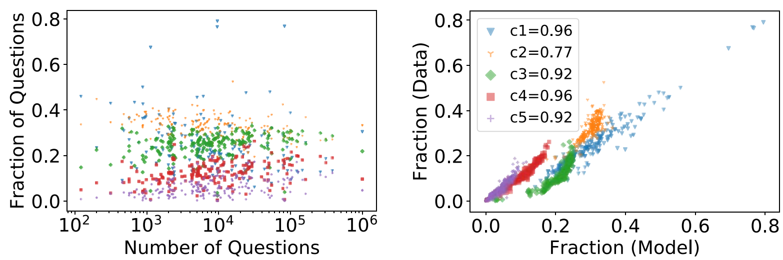

FIG. 6. Distributions of number of tags per post. Markers indicate number of tags: 1—blue triangle; 2—yellow 'Y'; 3—green diamond; 4—red square; and 5—purple '+'. (*Left*) Fraction of questions with a given number of tags as a function of the number of questions in a datasets. The distribution of tags per post is roughly independent of the size of the Stack Exchange community. (*Right*) Comparison of the distribution of the number of tags per post in the data and a sample from our generative model. The distributions are strongly correlated (shown in legend).

tags with our generative model. We generated tag-question bipartite graphs with Algorithm 1 for each Stack Exchange community using the fitted lognormal parameters (Fig. 4, bottom). The mean fraction of questions with more than five tags in the generated networks across 168 Stack Exchange platforms is only 2.5%, and furthermore, more than 80% of datasets have less than 4.5% of questions with more than five tags (Fig. 5, right).

4.3 *Summary*

Algorithm 1 is a simple generative model for bipartite tag-question networks that generates tag occurrences with the lognormal distribution that we found to be common across nearly all Stack Exchange communities. As a first look at how our model matches the empirical data, we consider the distribution of the number of tags per question. In the empirical data, this distribution tends to be uncorrelated with the size of the dataset (Fig. 6, left). We also find that the distribution of the number of tags per question in the model closely matches the empirical data (Fig. 6, right). In the next section, we analyse co-tagging, that is, how tags jointly annotate questions. Our model has no built-in notion of correlations in co-tagging, yet we find that the model still matches macroscopic co-tagging properties in the data.

5. Co-tagging analysis

In addition to the bipartite tag-question network, we also build a 'co-tagging network' for each Stack Exchange community. Recall that the tag-question network $B = (T \cup Q, E)$ is given by vertex sets T and Q corresponding to tags and questions and has edges $(t, q) \in E$ connecting tags to questions. The co-tagging network G is the projection of this graph onto the set of tags. Formally, $G = (T, F)$, where $(s, t) \in F$ if and only if there is some question $q \in Q$ such that $(s, q), (t, q) \in E$. In this case, we say that s and t *co-tag* with each other. We also associate a weight with each edge in G corresponding to the number of questions containing the two tags (the number of times that two nodes are co-tagged):

$$w_{s,t} = |\{q \in Q \mid (s, q), (t, q) \in E\}|. \tag{5.1}$$

In the rest of this section, we show that co-tagging networks constructed from samples of our generative model (Algorithm 1) match statistical properties of the co-tagging networks of empirical data, even though our model does not explicitly account for co-tagging behaviour. Again, we use the lognormal parameters μ and σ fitted for each dataset (Fig. 4) to generate a random graph for each Stack Exchange network. We focus our attention on three properties of the co-tagging network: (i) the expected number of co-tags (i.e. the weighted degree in G) as a function of tag frequency; (ii) the expected number of unique co-tags (i.e. the unweighted degree in G), again as a function of tag frequency; and (iii) weighted and unweighted versions of the clustering coefficient of the graph G.

5.1 *Weighted co-tags and tag frequency*

We first examine the relationship between the number of co-tags of a given tag as a function of its frequency (the number of questions in which it appears). We consider the number of co-tags to be weighted, that is, the number of co-tags of tag t is $k_t = \sum_{s \in T} w_{s,t}$, following Eq. (5.1). In the empirical data, this relationship is essentially linear. A linear model of the number of co-tags regressed on the number of questions containing the tag has a coefficient of determination (r^2 value) greater than 0.95 in 95% of the Stack Exchange communities. Figure 7 (left) shows the distribution of the slopes, which concentrates around 1.82.

We now show why we would also expect this behaviour from our model. Recall that the generative model samples tag frequencies $x_t \sim \mathrm{LogNormal}(\mu, \sigma^2)$ and then scales this sample so that these variables

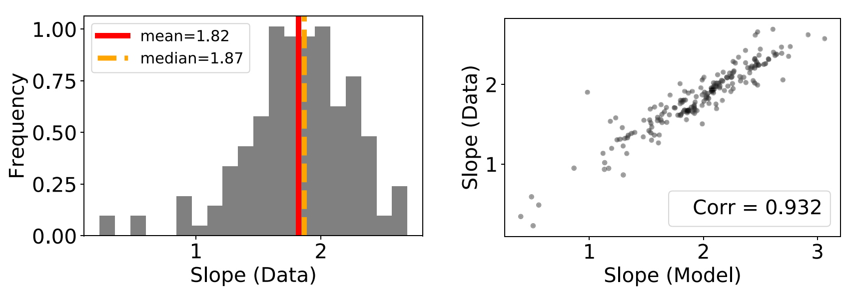

FIG. 7. (*Left*) The weighted number of co-tags is approximately a linear function of tag frequency. Here, we show the distribution of slopes from the linear regression over our collection of Stack Exchange communities. The regression has an r^2 value greater than 0.95 in 95% of the empirical datasets and greater than 0.97 in 97% of the generated datasets. (*Right*) The relationship between the fitted slope on the data and in the model across the Stack Exchange communities, which are strongly correlated.

match the total number of tag occurrences. The number of co-tags between s and t then follows $w_{s,t} \sim \text{Hypergeom}(\hat{N}_Q, x_s, x_t)$, where $\hat{N}_Q$ is the corrected number of questions in Algorithm 1. Thus, the expected number of co-tags k_t of a given tag t is

$$\mathbb{E}[k_t] = \sum_{t \neq s, s=1,\ldots,N_T} \frac{x_s x_t}{\hat{N}_Q} = \frac{(m - x_t)x_t}{\hat{N}_Q}, \tag{5.2}$$

where m is the target number of tag occurrences (the first equality comes from the independence in assignment of the tags). Although there is a quadratic relationship between k_t and x_t, we know that x_t is typically small compared to m. Thus, the gradient is well approximated by the linear function $m/\hat{N}_Q$, that is, $\frac{d}{dx_t} k_t \approx m/\hat{N}_Q$, independent of x_t. (Our analysis here is independent of the lognormal distribution of the tag frequency; we only relied on independence in the way that tags are assigned to questions.)

In actual random samples, the linear relationship holds. We performed the same linear regression on random samples from our generative model using the fitted parameters in Fig. 4 as we did for the empirical datasets. In the model, 97% of the 168 datasets have a correlation coefficient $r^2 > 0.97$. Furthermore, the slopes from the regression on the generated data are highly correlated with the slopes on the empirical data (the correlation is 0.932; Fig. 7, right), and the mean-squared error between the slope derived from a sample from the generative model and the computed slope on the empirical data across all Stack Exchange communities is just 0.10.

5.2 *Unique co-tags and tag frequency*

In the above analysis, we saw that the number of co-tags of a given tag is approximately linear in the number of questions in which the tag appears in both the empirical data and our model-generated data. In this section, we instead consider the number of *unique* co-tags of a given tag t as a function of the number of questions containing tag t. In this case, the number of unique co-tags is equal to the unweighted degree of tag t in the co-tagging network G defined above.

We find that the log of the number of unique co-tags is well approximated by a third-degree polynomial of the log of the number of question that contain the tag. Formally, let d_t denote the unweighted degree of tag t in the co-tagging network G and x_t the number of questions containing tag t. We then fit the following polynomial model:

$$\log(d(t) + 1) = \sum_{i=0}^{3} a_i \log(x_t + 1)^i. \tag{5.3}$$

Figure 8 (left) shows the CDF of the mean-squared error of the polynomial fit. The third-degree polynomial is a good fit for both the empirical data and the model across the collection of Stack Exchange communities. Figure 9 shows the distributions and fit of the third-degree polynomial for a few representative networks. In these cases, the polynomial fit is accurate and captures the fact that the number of unique co-tags does not grow linearly with tag frequency. Instead, the growth in unique co-tags tapers for the most frequently used tags. This happens because there is a limited total number of tags (Fig. 2), so tags that occur frequently have fewer options to increase the number of unique co-tags.

Interestingly, the fitted third-degree polynomial coefficients $\{a_i\}$, when taken as a collection across the Stack Exchange communities, largely lie on a lower-dimensional subspace. In the empirical datasets, the first principal component explains 86% of the variability, and the second principal component explains an

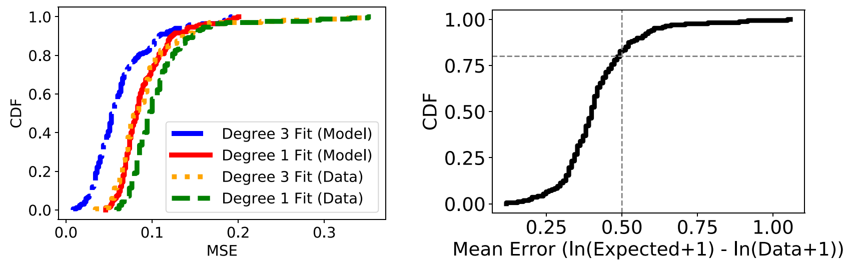

FIG. 8. (*Left*) CDF of the mean-squared error in third-degree and first-degree (linear) polynomial models of the log number of unique co-tags in terms of the log of tag frequency in both the data and the samples from the generative model. The third-degree polynomial is a good approximation and matches the expected value of the model (Fig. 9). (*Right*) CDF of the mean error in the expected number of unique co-tags in the model and the empirical number of co-tags. The error is less than 0.5 in 80% of the datasets. The model slightly over-estimates the number of unique co-tags by ignoring account tag correlations (see also Fig. 9).

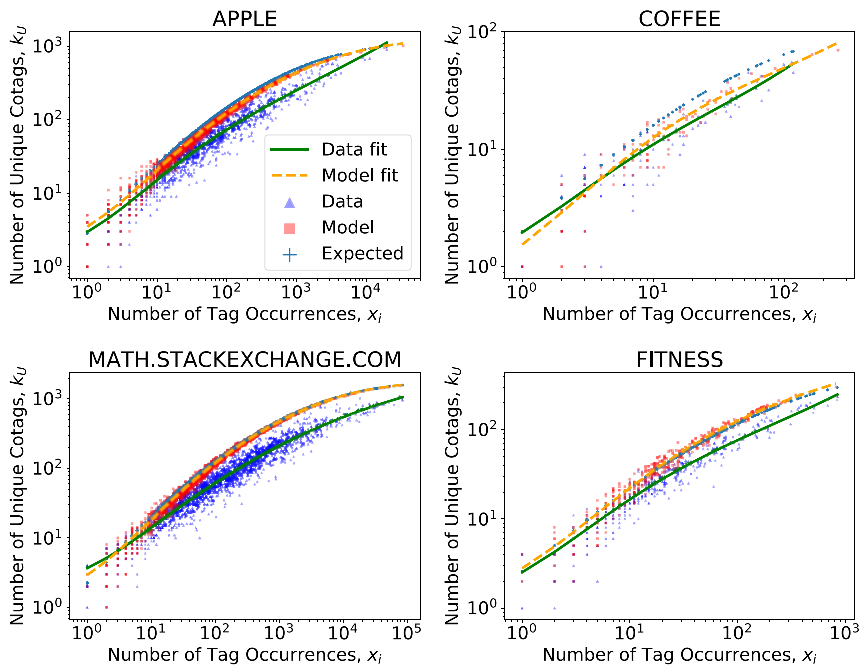

FIG. 9. Relationship between the number of unique co-tags and tag frequency on four Stack Exchange communities, which is well approximated by a degree-three polynomial (see also Fig. 8). The model has the same shape, albeit slightly above the data.

additional 13% of the variability. Similar results hold for the fitted coefficients in datasets generated with our model—89% of the variability is explained with the first principal component and an addition 10% is explained by the second principal component. Again, statistical properties of tagging persist across the collection of Stack Exchange communities, even though the communities themselves develop largely independently and cover a variety of topics. Our empirical findings suggest that tagging behaviour is based on the general Stack Exchange platform structure and not the individual communities' behaviours.

Finally, we show how to compute the expected number of unique co-tags in the model with a simple summation. We argued in the previous section that the weighted number of co-tags between tags s and t is $w_{s,t} \sim \text{Hypergeom}(\hat{N}_Q, x_s, x_t)$. Thus, the expected number of unique co-tags d_t of tag t is

$$\mathbb{E}[d_t] = \sum_{s \neq t} \mathbb{P}(w_{s,t} > 0) = \sum_{s \neq t} 1 - \mathbb{P}(w_{s,t} = 0) = \sum_{s \neq t} \left[1 - \binom{\hat{N}_Q - x_s}{x_t} \Big/ \binom{\hat{N}_Q}{x_t} \right], \tag{5.4}$$

where x_s is the sampled number of questions for tag s in Algorithm 1 and $\hat{N}_Q$ is the corrected number of questions. Figure 9 shows that the generated model data matches this expectation.

5.3 *Clustering in the co-tagging networks*

Finally, we analyse the clustering coefficient of the co-tagging networks, which is one of the fundamental measurements in networks [8, 36]. Let Δ_u, d_u and $w_{u,v}$ be the number of triangles containing node u, the unweighted degree of d_u, and the weight of edge (u, v) in the co-tagging graph G. We consider three clustering coefficients:

1. The unweighted clustering coefficient [36]:

$$C = \frac{1}{|T|} \sum_{u \in T} \frac{2\Delta_u}{d_u(d_u - 1)}. \tag{5.5}$$

2. The weighted clustering coefficient:

$$C_w = \frac{1}{|T|} \sum_{u \in T} \frac{1}{d_u(d_u - 1)} \sum_{v,z} (\hat{w}_{u,v} \hat{w}_{u,z} \hat{w}_{v,z})^{1/3}, \tag{5.6}$$

   where $\hat{w}_{u,v} = w_{u,v} / \max_{x,y} w_{x,y}$ [37]. We will analyse $\log(C_w)$.

3. The log-weighted clustering coefficient, which is the same as the mean weighted clustering coefficient, except the weight $w_{u,v}$ is replaced by $w'_{u,v} = \log(w_{u,v} + 1)$:

$$C_{lw} = \frac{1}{|T|} \sum_{u \in T}^{n} \frac{1}{d_u(d_u - 1)} \sum_{v,z} (\hat{w}'_{u,v} \hat{w}'_{u,z} \hat{w}'_{v,z})^{1/3}, \tag{5.7}$$

   where $\hat{w}'_{u,v} = w'_{u,v} / \max_{x,y} w'_{x,y}$ and summations over cases where $w'_{u,v} = 0$ (i.e. with no edge) are ignored.

Figure 10 (top row) shows that all three clustering coefficients are approximately normally distributed across the collection of Stack Exchange communities. Furthermore, the unweighted coefficients are only weakly correlated with the size of the community, measured by the log-number of questions on the Stack

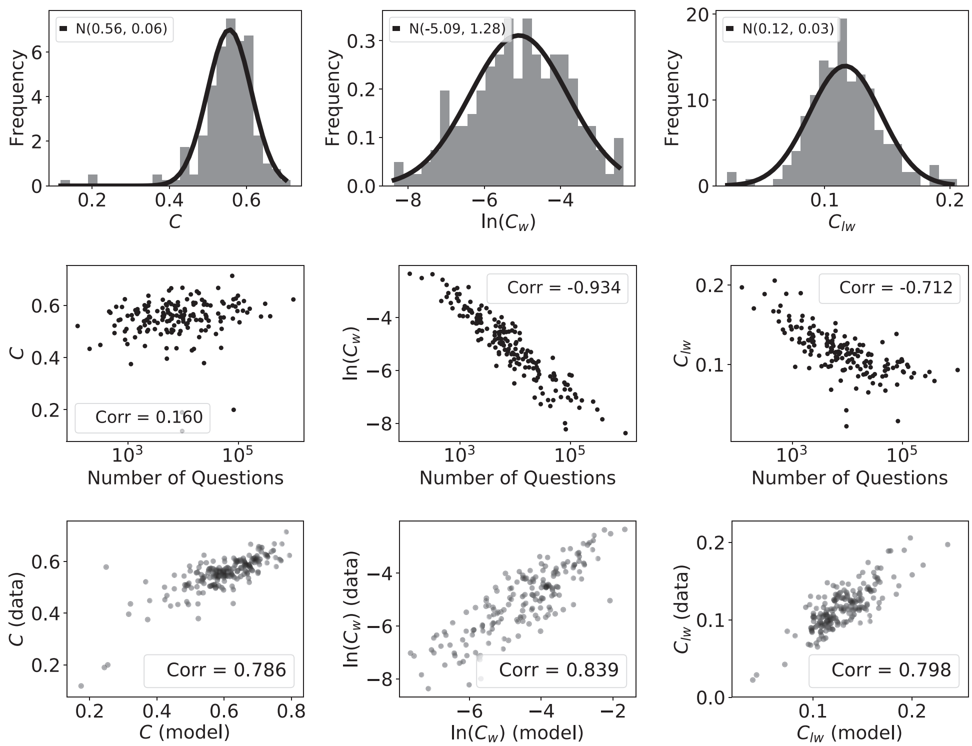

FIG. 10. (*Top*) The clustering coefficients of the Stack Exchange communities are approximately normally distributed. (*Middle*) The unweighted clustering coefficient C has a weak correlation with respect to the size of the community as measured by the log-number of questions; the weighted versions are negative correlated. (*Bottom*) The clustering coefficients in the co-tagging networks generated by our model are similar to the clustering coefficients of the empirical Stack Exchange communities.

Exchange (Fig. 10, middle row). We conclude that the size of a Stack Exchange community is likely not a driving factor in the unweighted clustering of the network, which backs up conventional wisdom for the analysis of real-world networks [8]; however, this differs from the behaviour of random graph models that produce heavy-tailed degree distributions, where clustering decreases with size [38] (Fig. 10, middle row). On the other hand, the weighted clustering coefficients tend to decrease with the size of the Stack Exchange community.

The co-tagging networks derived from samples of our generative model reproduce these clustering coefficients remarkably closely and with strong positive correlations (Fig. 10, bottom row). Again, we emphasize that our model does not bake in any explicit notion of clustering. Instead, our model only matches the lognormal distribution of the tag frequency and the total number of tags applied to all questions. Thus, clustering in the co-tagging in Stack Exchange communities could be explained simply by these simpler statistics. This finding contrasts sharply with typical (social) network analysis, where clustering is exhibited at a much higher level than is expected by random graph models [8]. The key difference is that our model is based on a projection of a bipartite tag-question graph rather than directly modelling the co-tagging network. This type of modelling has a long history in sociology [39] but has received relatively less theoretical attention in social network analysis [29].

6. Discussion

In addition to providing answers to questions, the users of Q&A platforms create knowledge through annotation of questions. With its tagging system, Stack Exchange provides a unique opportunity to study these annotations for two main reasons. First, tags cannot be created freely and there are community guidelines for their application, which differs substantially from tagging norms on other social media platforms. Second, there is a collection of Stack Exchange communities that have largely evolved independently, enabling us to model and analyse tagging with more statistical evidence. And we indeed found strong similarities in macroscopic tagging structure—in terms of tag frequency and co-tagging network structure—across 168 Stack Exchange communities spanning a diverse range of topics. This contrasts from typical network analyses that study a single snapshot of a social network. Previously, researchers have circumvented this issue by looking at, for example, sets of disparate subgraphs from a larger graph [40, 41]; samples of ego networks [42–44]; and collections of snapshots of time-evolving networks [45].

We used the tag frequency distribution to develop a simple generative model for random tag-question bipartite graphs, which was able to reproduce a number of the co-tagging and clustering properties of the datasets, without explicitly modelling correlations or clustering in the co-tagging process. Further understanding of the process producing this distribution is an avenue for future research. For example, multiplicative growth models are a well-known generative process for lognormal distributions [34]. Although outside the scope of this article, the availability of temporal information from Stack Exchange provides a path towards more robust understanding of the underlying processes of tag use, similar to other methods for estimating growth on the Web and in social networks [46, 47].

It has long been known to the network community that heavy-tailed degree distribution are common in various networks. Here, we choose the lognormal distribution as it offers a statistically better fit than other heavy-tail distributions for most of the 168 datasets according to our empirical analysis. We also investigated the distributional information of the fitted parameters of the 168 datasets. A third-degree polynomial model of co-tagging frequencies as a function of tag frequencies consistently gives good results on most of the datasets. Because the third-degree polynomial model fits the datasets well, again we can study the model parameters as a high level description of all the different datasets and analyse their distributions. This is a novel perspective to look at tagging in Q&A networks and shows a new discovery on them: despite entirely different topics, the tagging behaviour of each Q&A networks are strikingly similar. This results cannot be derived from other Q&A networks because no other popular Q&A networks have so many largely independently evolving communities focusing on such a variety of topics as Stack Exchange. Still, an interesting direction for future research would be to see if certain sub-communities within a single Q&A platform also exhibit similar structure. Finally, our results suggest that social tagging behaviour on Stack Exchange is more a product of the platform at large, as opposed to being driven by the individual Stack Exchange communities.

Code and data

Code to reproduce our results, along with processed data, are available at https://github.com/yushangdi/stack-exchange-cotagging.

Funding

National Science Foundation (NSF) (award DMS-1830274); Army Research Office (ARO) (award W911NF-19-1-0057 and MURI).

References

1. Wang, G., Gill, K., Mohanlal, M., Zheng, H. & Zhao, B. Y. (2013) Wisdom in the social crowd: an analysis of Quora. *Proceedings of the 22nd International Conference on World Wide Web (WWW). WWW'13, Rio de Janeiro, Brazil*. New York, NY, USA: ACM, pp. 1341–1352.

2. Adamic, L. A., Zhang, J., Bakshy, E. & Ackerman, M. S. (2008) Knowledge sharing and Yahoo answers: everyone knows something. *Proceedings of the 17th International Conference on World Wide Web (WWW). WWW'08, Beijing, China*. New York, NY, USA: ACM, pp. 665–674.

3. Chen, Y., Ho, T.-H. & Kim, Y.-M. (2010) Knowledge market design: a field experiment at Google answers. *J. Public Econ. Theory*, **12**, 641–664.

4. Anderson, A., Huttenlocher, D., Kleinberg, J. & Leskovec, J. (2013) Steering user behavior with badges. *Proceedings of the 22nd International Conference on World Wide Web (WWW). WWW'13, Rio de Janeiro, Brazil*. New York, NY, USA: ACM.

5. Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. (2011) The anatomy of the Facebook social graph. *CoRR*, **abs/1111.4503**. http://dblp.uni-trier.de/db/journals/corr/corr1111.html#abs-1111-4503.

6. Kwak, H., Lee, C., Park, H. & Moon, S. (2010) What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web (WWW). WWW'10, Raleigh, NC, USA*. New York, NY, USA: ACM.

7. Salganik, M. J. (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, **311**, 854–856.

8. Newman, M. E. J. (2003) The structure and function of complex networks. *SIAM Rev.*, **45**, 167–256.

9. Bosu, A., Corley, C. S., Heaton, D., Chatterji, D., Carver, J. C. & Kraft, N. A. (2013) Building reputation in StackOverflow: an empirical investigation. *10th Working Conference on Mining Software Repositories (MSR), San Francisco, CA, USA*. Piscataway, NJ, USA: IEEE, pp. 89–92.

10. Paul, S. A., Hong, L. & Chi, E. H. (2012) Who is authoritative? Understanding reputation mechanisms in Quora. *Collective Intelligence*. arXiv preprint arXiv:1204.3724.

11. Anderson, A., Huttenlocher, D., Kleinberg, J. & Leskovec, J. (2012) Discovering value from community activity on focused question answering sites: a case study of Stack Overflow. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). KDD'12, Beijing, China*. New York, NY, USA: ACM.

12. Posnett, D., Warburg, E., Devanbu, P. & Filkov, V. (2012) Mining Stack Exchange: expertise is evident from initial contributions. *International Conference on Social Informatics. Lausanne, Switzerland*. Piscataway, NJ, USA: IEEE, pp. 199–204.

13. Paranjape, A., Benson, A. R. & Leskovec, J. (2017) Motifs in temporal networks. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. *WSDM'17, Cambridge, UK*. New York, NY, USA: ACM.

14. Yang, J., Morris, M. R., Teevan, J., Adamic, L. A. & Ackerman, M. S. (2011) Culture matters: a survey study of social Q&A behavior. *International AAAI Conference on Web and Social Media (ICWSM)*, *Barcelona, Spain*. Palo Alto, CA, USA: AAAI Press.

15. Tian, Q., Zhang, P. & Li, B. (2013) Towards predicting the best answers in community-based Question-Answering services. *International AAAI Conference on Web and Social Media (ICWSM)*, *Boston, USA*. Palo Alto, CA, USA: AAAI Press.

16. Maity, S., Sahni, J. S. S. & Mukherjee, A. (2015) Analysis and prediction of question topic popularity in community Q&A sites: a case study of Quora. *International AAAI Conference on Web and Social Media (ICWSM)*, *Oxford, UK*. Palo Alto, CA, USA: AAAI Press.

17. MacLeod, L. (2014) Reputation on Stack Exchange: tag, you're it! *28th International Conference on Advanced Information Networking and Applications Workshops (AINA), Los Alamitos, CA*. Piscataway, NJ, USA: IEEE, pp. 670–674.

18. Pal, A., Chang, S. & Konstan, J. A. (2012) Evolution of experts in question answering communities. *International AAAI Conference on Web and Social Media (ICWSM), Dublin, Ireland*. Palo Alto, CA, USA: AAAI Press.

19. Weller, K., Bruns, A., Burgess, J., Mahrt, M. & Puschmann, C. (2014) *Twitter and Society*, vol. 89. Switzerland: Peter Lang International Academic Publishers.

20. An, J. & Weber, I. (2016) #greysanatomy vs. #yankees: demographics and Hashtag use on Twitter. *Tenth International AAAI Conference on Web and Social Media (ICWSM2016), Cologne, Germany*. Palo Alto, CA, USA: AAAI Press.

21. Vander Wal, T. (2005) Folksonomy. http://www.vanderwal.net/random/entrysel.php?blog=1622.

22. Capocci, A. & Caldarelli, G. (2008) Folksonomies and clustering in the collaborative system CiteULike. *J. Phys. A*, **41**, 224016.

23. Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G. & Loreto, V. (2009) Collective dynamics of social annotation. *Proc. Natl. Acad. Sci. USA*, **106**, 10511–10515.

24. Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V. D. P., Loreto, V., Hotho, A., Grahl, M. & Stumme, G. (2007) Network properties of Folksonomies. *AI Commun.*, **20**, 245–262.

25. Bascompte, J., Jordano, P. & Olesen, J. M. (2006) Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science*, **312**, 431–433.

26. Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. & Barabasi, A.-L. (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.

27. Akoglu, L., Chandy, R. & Faloutsos, C. (2013) Opinion fraud detection in online reviews by network effects. *International AAAI Conference on Web and Social Media (ICWSM) Boston, USA*, vol. 13. Palo Alto, CA, USA: AAAI Press, pp. 2–11.

28. Larremore, D. B., Clauset, A. & Jacobs, A. Z. (2014) Efficiently inferring community structure in bipartite networks. *Phys. Rev. E*, **90**, 012805.

29. Lattanzi, S. & Sivakumar, D. (2009) Affiliation networks. *Proceedings of the 41st annual ACM Symposium on Theory of Computing (STOC), Bethesda, MD, USA*. New York, NY, USA: ACM.

30. Chojnacki, S. & Kłopotek, M. (2010) Random graph generative model for Folksonomy network structure approximation. *Procedia Comput. Sci.*, **1**, 1683–1688.

31. Halpin, H., Robu, V. & Shepherd, H. (2007) The complex dynamics of collaborative tagging. *Proceedings of the 16th International Conference on World Wide Web (WWW). WWW'07, Banff, Alberta, Canada*. New York, NY, USA: ACM.

32. Wang, X., Liu, H. & Fan, W. (2011) Connecting users with similar interests via tag network inference. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11), Glasgow, Scotland, UK*. New York, NY, USA: ACM, pp. 1019–1024.

33. Clauset, A., Shalizi, C. & Newman, M. (2009) Power-law distributions in empirical data. *SIAM Rev.*, **51**, 661–703.

34. Mitzenmacher, M. (2004) A brief history of generative models for power law and lognormal distributions. *Internet Math.*, **1**, 226–251.

35. Alstott, J., Bullmore, E. & Plenz, D. (2014) Powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one*, **9**, e85777–e85777.

36. Watts, D. J. & Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

37. Onnela, J. P., Saramki, J., Kertsz, J. & Kaski, K. (2005) Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E.*, **71**, 065103.

38. Bollobás, B. & Riordan, O. M. (2004) Mathematical results on scale-free random graphs. *Handbook of Graphs and Networks*. New York, NY, USA: John Wiley & Sons, Inc., pp. 1–34.

39. Breiger, R. L. (1974) The duality of persons and groups. *Soc. Forces*, **53**, 181–190.

40. Patania, A., Petri, G. & Vaccarino, F. (2017) The shape of collaborations. *EPJ Data Sci.*, **6**, 18.

41. Traud, A. L., Mucha, P. J. & Porter, M. A. (2012) Social structure of Facebook networks. *Phys. A*, **391**, 4165–4180.

42. Benson, A. R., Abebe, R., Schaub, M. T., Jadbabaie, A. & Kleinberg, J. (2018) Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. USA*, **115**, E11221–E11230.

43. McAuley, J. & Leskovec, J. (2014) Discovering social circles in ego networks. *ACM Trans. Knowl. Discov. Data*, **8**, 1–28.

44. Ugander, J., Backstrom, L. & Kleinberg, J. (2013) Subgraph frequencies: mapping the empirical and extremal geography of large graph collections. *Proceedings of the 22nd International Conference on World Wide Web (WWW) Rio de Janeiro, Brazil*. New York, NY, USA: ACM.

45. Yaveroğlu, O. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A. & Pržulj, N. (2014) Revealing the hidden language of complex networks. *Sci. Rep.*, **4**, 4547.

46. Huberman, B. A. & Adamic, L. A. (1999) Growth dynamics of the World-Wide Web. *Nature*, **401**, 131–131.

47. Overgoor, J., Benson, A. R. & Ugander, J. (2018) Choosing to grow a graph: modeling network formation as discrete choice. *The World Wide Web Conference. WWW'19, San Francisco, CA, USA*. New York, NY, USA: ACM, pp. 1409–1420.