Discovery of Topical Authorities in Instagram

Aditya Pal, Amaç Herdağdelen, Sourav Chatterji, Sumit Taank, Deepayan Chakrabarti*
Facebook
{apal,amac,sourav,staank}@fb.com, deepay@utexas.edu

ABSTRACT

Instagram has more than 400 million monthly active accounts who share more than 80 million pictures and videos daily. This large volume of user-generated content is the application's notable strength, but also makes the problem of finding the authoritative users for a given topic challenging. Discovering topical authorities can be useful for providing relevant recommendations to the users. In addition, it can aid in building a catalog of topics and top topical authorities in order to engage new users, and hence provide a solution to the *cold-start* problem.

In this paper, we present a novel approach that we call the Authority Learning Framework (ALF) to find topical authorities in Instagram. ALF is based on the self-described interests of the follower base of popular accounts. We infer regular users' interests from their self-reported biographies that are publicly available and use Wikipedia pages to ground these interests as fine-grained, disambiguated concepts. We propose a generalized label propagation algorithm to propagate the interests over the follower graph to the popular accounts. We show that even if biography-based interests are sparse at an individual user level they provide strong signals to infer the topical authorities and let us obtain a high precision authority list per topic. Our experiments demonstrate that ALF performs significantly better at user recommendation task compared to fine-tuned and competitive methods, via controlled experiments, in-the-wild tests, and over an expert-curated list of topical authorities.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, retrieval models, selection process*; H.1.2 [**Information Systems**]: User/Machine Systems—*human factors, human information processing*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Topical Authorities, User Recommendation, Instagram

*Author has relocated to McCombs School of Business, University of Texas, Austin, TX, USA.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883078.

1. INTRODUCTION

Instagram is one of the most popular online photo and video sharing services, having more than 400 million active accounts per month who in turn share more than 80 million photos and videos per day. This large volume of user-generated content leads to a rich diversity of topics on Instagram and one can find high quality pictures on even niche topics: e.g., one can browse pictures on *origami* at `https://instagram.com/explore/tags/origami`. However finding *users* that specialize in the topic *origami* can be quite challenging. Discovering the topically sought people (topical authorities) can help in providing relevant recommendations to the users. In addition, it can aid in building a catalog of topics and authorities in order to engage new users.

There has been a considerable effort towards authority discovery in several domains, such as microblogs [37, 31, 16], emails [10], community question answering [28, 38, 32], and enterprise corpora [3, 30]. However, Instagram poses three unique challenges primarily due to the nature of content shared on it, how users interact with it, and also in part due to our problem specification. We highlight these challenges below and also provide an intuitive reasoning as to why most prior work does not perform as well in our setting.

Sparsity of Textual Features. Unlike other social media domains, Instagram has rich visual content but terse textual information. Hence, algorithms that depend on the text-based user features would not perform well for this domain. Prior work [36] also highlights this issue in other social media domains that content-boosted methods can suffer due to brevity and sparseness of text documents.

Misleading Topic Signals from Users' Activity. Users (especially celebrities who are authorities on a specific subject) typically share general posts like pictures of their family & friends, events they have attended, products they have bought, causes they are concerned about, etc. Table 1 shows the relative probability of the most used hashtags by a group of well-known basketball players. We note here that only the hashtag '*theland*' is somewhat related to basketball while the rest are too generic. Algorithms that infer authority based on users' activity would either ignore these players or assign

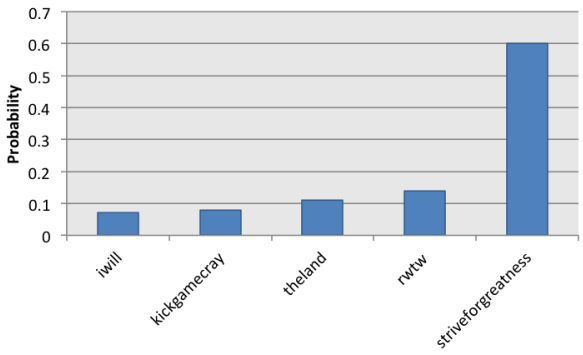

Figure 1: Normalized frequency of the five most used hashtags by a group of well-known basketball players in one month period.

them generic topics; thus they would not be recommended to users interested in basketball.

Interpretability of Topics. For new users, recommendation algorithms can typically only provide generic recommendations as they have little to no signal about them (also referred to as the *cold-start* problem). Hence it would be useful to present a catalog of curated topics and top authorities within those topics for use by new users. Popular topic models, such as LSI [11], pLSI [19], and LDA [6] define topics in high dimensional word space. These embeddings can merge several related concepts together and mask their relative importance. E.g., consider a topic embedding that merges concepts, such as *dog* and *cat*. Since dogs and cats are quite popular on Instagram, it would be more useful to show them as separate topics than together. In most cases, *niche* topics get buried in their popular counterparts, like *origami* within *art*. At the very least, the merged concepts can be confusing to a new user.

Most prior models on topical authority discovery do not perform as well in the context of Instagram due to one or more of the above-mentioned challenges. Models that are based on users' contribution fail to perform well due to the first and second issue. Graph-based [8, 25] and collaborative filtering [18, 29, 26, 24] models are less sensitive to users' content, but do not provide an explicit set of topics that can be readily shown to new users. Furthermore these models can be crude in their recommendations as they look at user-user similarity for recommendations and can miss out the niche interests of users that distinguish them from other users. Some models, such as TwitterRank [37], require the construction of a topically weighted graph which can be problematic due to misleading topic signals from users' activity. Moreover as prior work [31] suggests that these methods can be prone to surfacing celebrities as topical authorities.

We propose an authority learning framework (ALF) that side-steps these issues through the following design choices.

Topic vocabulary from Wikipedia. Wikipedia pages are well defined; hence topics based on them can be incorporated to build a topic catalog. For instance, *hachimaki*[1] can be a topic, albeit niche; unlike latent topic approaches, we do not merge it with a relevant but popular topic such as *clothing accessories*. This is a crucial first step as there might be niche audience for these niche topics and our goal is to cater to their needs. Additionally, *Wikified* topics simplifies the task of ground-truth collection and validation; e.g., it is trivial to verify the assignment of an *NBA player* to *basketball* than to a latent topic vector.

Infer interests from users' biographies. Instagram users can fill out a publicly viewable field called *biography description* where they can provide free-text about themselves. Among other things, they may choose to share their profession, interests, etc. This is a sparse feature for individual users because not everyone provides a *publicly-available* description and neither does a user specify all her interests in this section. However, when aggregated among followers of popular accounts, they can provide meaningful information about the account being followed.

Estimate authorities from followers' interests. We hypothesize that "*an authority on a specific topic has a significantly higher proportion of followers interested in that topic*". We operationalize this hypothesis by proposing a generalized label propagation algorithm that propagates the user interests over the follower graph. Our algorithm is a generalization of the label propagation algorithm as it handles the scenario where only positive (or negative) labels are present in the graph. Additionally, it allows us to trade between the "*explainable*" and "*broader*" inferences depending on the business needs. Finally, we compute the authority scores from the label scores through a topic specific normalization and processing of the false-positives. We note that while several graph-based approaches such as PageRank [8] (and its variants) nominally employ a similar hypothesis, however their direct application to our problem does not yield accurate results, as we show experimentally.

Our approach is designed to handle the scale of data at Instagram and it is tailored to have high precision while still being computationally efficient. We conduct controlled experiments, in-the-wild tests, and over an expert-curated list of topical authorities to show the effectiveness of the proposed method in comparison to fine-tuned and competitive prior methods. Our method yields over 16% better click-through and 11% better conversion rates for user recommendation task than the closest alternative method, and a qualitative analysis of $24,000$ (authority, topic) assignments by ALF were judged to have a precision of 94%.

Outline: The rest of the paper is organized as follows. We discuss the related work in Section 2. We describe our design decisions in Section 3 and formally introduce our model in Section 4. Section 5 outlines the real-time recommender based on the output of our model. Experimental evaluation is discussed in Section 6, followed by conclusion in Section 7. Proofs are deferred to the appendix.

2. RELATED WORK

Finding authoritative users in online services is a widely studied problem. We discuss some popular methods and application domains next.

Graph-based approaches. Among the most popular graph based algorithms are PageRank [8], HITS [25] and their variants, such as authority rank [13] that combines social and textual authority through the HITS algorithm for the World Wide Web (see [7] for a comprehensive survey). While graph-based ranking algorithms such as PageRank and HITS

[1]https://en.wikipedia.org/wiki/Hachimaki

(on topically weighted graphs) are very popular, they do not work well in our context because they are prone to surfacing celebrities since their repeated iterations tend to transfer weight to the highly connected nodes in the graph. We solve this issue by proposing a generalized label propagation algorithm that enables us to control for scores that are easily explainable (i.e. from graph neighbors) and broader (i.e. transferred over a path in the graph). Unlike PageRank, the label propagation algorithm essentially penalizes users that do not have a very topically specific following, which deters overly general celebrities from dominating the authority lists. However that alone is not sufficient; we also show how label scores can be used to generate users' authority scores through a topic specific normalization and a series of post-processing steps, such as false positive removal, to obtain a high quality list of authorities.

E-mail and Usenet. Fisher et al. [14] analyzed Usenet newsgroups which revealed the presence of "answer people", i.e. users with high out-degree and low in-degree who reply to many but are rarely replied to, who provide most answers to the questions in the community. Campbell et al. [10] used a HITS-based graph algorithm to analyze the email networks and showed that it performed better than other graph algorithms for expertise computation. Several efforts have also attempted to surface authoritative bloggers. Java et al. [20], applying models proposed by Kempe et al. [23], model the spread of influence on the Blogosphere in order to select an influential set of bloggers who maximize the spread of information on the blogosphere.

Question Answering. Authority identification has also been explored extensively in the domain of community question answering (CQA). Agichtein et al. [1] extracted graph features such as the degree distribution of users and their PageRank, hubs and authority scores from the Yahoo Answers dataset to model a user's relative importance based on their network ties. They also consider the text based features of the question and answers using a language model. Zhang et al. [38] modified PageRank to consider whom a person answered in addition to how many people a person answered. They combined the number of answers and number of questions of a user in one score, such that higher the score higher the expertise. Jurczyk et al. [22] identified authorities in Q&A communities using link analysis by considering the induced graph from interactions between users.

Microblogs. In the microblog domain, Weng et al. [37] modeled Twitter in the form of a weighted directed topical graph. They use topical tweets posted by a user to estimate the topical distribution of the user and construct a separate graph for each topic. The weights between two users indicate the degree of correlation between them in the context of the given topic. A variant of PageRank called TwitterRank is run over these graphs to estimate the topical importance of each user. Pal et al. [31] proposed a feature-based algorithm for finding topical authorities in microblogs. They used features to capture users' topical signal and topical prominence and ran clustering algorithms to find clusters of experts; these users are then ranked using a Gaussian-based ranking algorithm.

Recently, Popescu et al. [33] proposed an expertise modeling algorithm for Pinterest. They proposed several features based on users contributions and graph influence. Pinterest allows users to share categories along side their content and users were ranked based on their category-based activity.

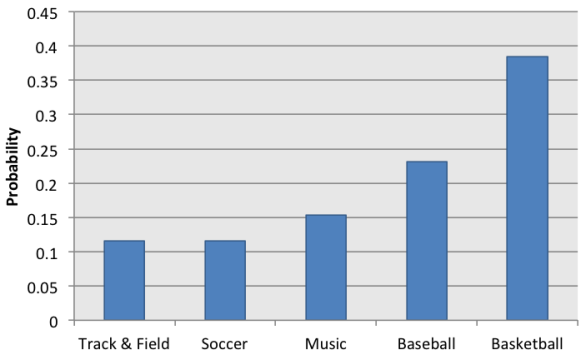

Figure 2: The fraction of topical followers to the total number of followers for the set of basketball players that were selected in the example of Fig. 1. *Only the top 5 topics based on fractions are selected and the fractions are then normalized to sum to 1.*

In summary, the notion of finding authorities has been explored extensively in other domains and has been dominated by network analysis approaches, often in conjunction with textual analysis. Within the photo-sharing arena, there is relatively little work on the issue of authority identification with the notable exception of [33]. Our model extends research in the authority detection arena by bringing a fresh perspective in modeling users' interests through their biographies and computing topical expertise through the label propagation of followers' interests. We propose a series of steps such as topic-based normalization and elimination of false positives to obtain highly accurate set of topical authorities. Our approach is computationally efficient and is designed to handle the scale at Instagram.

3. DESIGN CHOICES

We begin by exploring several design choices and assumptions that are fundamental to our model. We intuitively show that these choices lead to an accurate representation of authority among Instagram users.

3.1 Authority via Follower Interests

The first design question is: *what is an effective authority signal* in the context of Instagram? Conventional methods of authority in social media that rely upon textual features (such as [28, 38, 31]) do not work here because users can post overly general pictures with little to no textual information in them. This phenomenon is highlighted in the example of basketball players (Fig. 1).

On the other hand, if we examine the fraction of topical followers[2] of these basketball players (see Fig. 2), we observe that *basketball* surfaces at the top. Based on the proportions of basketball followers, these players lie within the $90-100$ percentile across all popular users – a strong reflector of their basketball prowess. This leads to the following hypothesis.

[2]Number of followers interested in a given topic by the total number of followers. The precise definition of topical interest will be presented later.

	Hashtag	Coverage	Tf-Idf
Trending	tbt	*1*	*1*
	wcw	0.52	0.70
	potd	0.51	3.46
	like4like	0.36	5.02
	tagsforlikes	0.30	4.80
Social	love	0.97	2.47
	family	0.56	0.78
	bff	0.21	0.72
	friend	0.12	1.28
Topical	fashion	0.50	3.07
	gym	0.22	3.15
	baseball	0.08	1.06
	basketball	0.07	1.45
	technology	0.02	1.58

Table 1: Coverage (% of population using the hashtag) and tfidf of different hashtags based on one month consumption data. *Here the statistics are divided with the statistics of the hashtag tbt for a relative comparison.*

HYPOTHESIS 1. *An authority on topic t has a significantly higher proportion of followers interested in t.*

This hypothesis is central to our approach, and is employed by several popular and successful graph based algorithms as well [8, 25, 13, 37].

3.2 Interests via User Biographies

The underlying assumption of the previous choice is that we are able to identify and extract users interests. Clearly we cannot use the "produced" content for this purpose. Alternately, one can consider the content consumed by the user (liked and/or commented). Yet the consumed content can lead to misleading interests, due to following issues:

I1 *Not all users login regularly and consume content.* Sporadic activity patterns can result in sparsity in interest estimation – undermining authority estimation. E.g., if we sorted all users according to the fraction of their followers who consumed the hashtag *basketball* over one-month period, the basketball players (from our example in Fig. 1) would only fall in the $80-90$ percentile among all popular users. This is a considerable underestimate of their authority on basketball.

I2 *Trending or agenda-driven topics can mask core interests.* Daily or weekly trending topics, such as *throw back Thursday* (*tbt*), *women crush Wednesday* (*wcw*), *photo of the day* (*potd*) engage a large set of users regularly. Moreover, several content producers use special hashtags (e.g. *like4like*, *follow4follow*, *tags4likes*) to communicate with their audience - eliciting an action from them. Table 1 shows the statistics of these different types of hashtags indicating that some of the non-topical hashtags can overwhelm in terms of their popularity and yet be competitive on their *tf-idf* scores.

I3 *Friends and family effects.* Due to social nature of Instagram most users follow their friends and hence their activity is mired with casual likes and comments on their friends posts. As a result, hashtags like *love*, *family*, *friend* appear as potential interests in Table 1.

I1-3 adds sparsity and noise to the interest estimation. We sidestep these issues by considering the self reported biographies of users. Extraction of interests from user biographies has also been explored by prior work (see for example [12]) and it offers several advantages: (1) users do not change their biographies frequently, and (2) they are independent of login/activity patterns. These two aspects make interest inference less sensitive to *trending*, *agenda*, *social*, and *spam* topics – providing a relatively noise-free set of interests. Biographies also address the coverage issue to an extent, since many users have publicly available non-empty biographies. We now make the following observation:

OBSERVATION 1. *Users tend to follow at least some accounts that match the interests reported in their biographies.*

Observation 1 in conjunction with Hypothesis 1 is akin to the concept of preferential attachment [4] along the topical lines. Intuitively it makes sense for observation 1 to hold for most users. For users where it does not hold, there is a clear opportunity for recommendation algorithms to fill the gap.

3.3 Scope of Topics

Our next design choice pertains to defining the scope of interests (topics) extracted from the biographies. Popular topic models, such as LSI [11], pLSI [19], and LDA [6] define topics to be embeddings in a high dimensional word space. However, these embedding are hard to interpret and label. From the point of view of a topic catalog, these topic embeddings cannot be directly shown to an end-user, as they can confusingly merge several concepts together. Moreover, the biographies from which we want to extract interests are short text mired with typos and abbreviations, rendering embedding formed from biographical text less useful. Finally, our choice of topics must also take into consideration the following aspects:

- *Treating correlated topics separately.* In context of Instagram, topics such as *nature*, *earth*, *flower*, *plants* can be highly correlated. Merging these seemingly related topics would be non-desirable for end users with finer tastes and also for content producers that focus on a niche topic.
- *A topic can be annotated by different words.* For example, both '*lakers*' and '*l. a. lakers*' point towards the basketball team '*Los Angeles Lakers*'. We must ensure that such annotations are merged in the canonicalized representation of the topic.

We handle above aspects by scoping a canonical topic to be one having a Wikipedia page. There are several advantages of this choice: (1) it implicitly respects the topic correlations as most nuanced topics have dedicated Wikipedia pages, (2) it provides a canonical representation for a topic, which makes it easier to identify the different variations of that topic, and (3) Wikipedia categories can be used for blacklisting or whitelisting certain types of topics.

Unlike embedded topics, our topics can be utilized to *explain* recommendations, such as, "if u follows x, and x is an authority on t, then u might be interested in t". Formally,

$$(u \rightarrow x) \wedge Authority(x,t) \Rightarrow Interest(u,t). \quad (1)$$

It is easier to verify the above claim manually than a similar claim over a latent topic vector.

3.4 One-to-One Authority Topic Mapping

We make a key design choice of restricting a user to be an authority on at the most one topic. Formally,

$$Authority(u,t) \Rightarrow \nexists t'(t' \neq t) \wedge Authority(u,t'). \quad (2)$$

From a practical point of view, this choice is necessary to restrict popular users from dominating several topics at once. In Fig. 2, we observe that the selected basketball players have a high probability score for topic *baseball* as well. If the one-topic restriction is not enforced, they would appear as an authority on *baseball* along side *basketball* - a scenario we wish to avoid. While there can be instances where a user dabbles in multiple topics (perhaps due to close relations between those topics), our restriction would surface that user as an authority on only one of the topics. We consider this acceptable since precision of authority detection is key; we are tolerant to a partial authority representation but not an inaccurate one. *Note, however, that a user is allowed to have multiple interests (only authority assignments are restricted).*

4. AUTHORITY LEARNING FRAMEWORK

The complexity of the problem precludes a simple global objective function that can be optimized to yield the authority scores. Instead, we propose to split the problem into three well-defined stages, each of which can be individually refined and tested. Fig. 3 presents the high level overview of our authority learning framework (ALF). The first step is the high-precision inference of the topical interests of users from their publicly available biographies. As the figure suggests, we infer from user A's biography that she is interested in topic t_1. Note that interests are inferred for only those users who have filled in their biography section. The next step is the joint inference of interests of all users, along with baseline authority scores, via propagation of the interests over the follower graph. For this purpose, we propose a generalized label propagation algorithm and present a practical instantiation of this algorithm that is easy to implement and parallelize. Finally, authority topics are assigned to the users through normalization and post-processing on the authority scores (user B is assigned topic t_1).

Notations. Formally, we have the follower graph $\mathcal{G} = (V, E)$ with V representing all Instagram users and edge $(u \rightarrow v) \in E$ indicate that user u follows user v. Let $n_v^{in} = |(u \rightarrow v) \in E|$ be the number of incoming edges to v and $n_v^{out} = |(v \rightarrow u) \in E|$ be the number of outgoing edges from v. Let $\mathcal{T}$ indicate the set of topics and $I(u) \subseteq \mathcal{T}$ indicate the topical interests extracted from u's self-reported biography.

4.1 Topic Vocabulary & User Interests

From a large list of top-level Wikipedia categories, an expert curator whitelisted a subset after filtering out categories that were irrelevant for our problem (e.g., *organizations*, *players*, *religion*, *locations*, *books*, *languages*, etc.). We

Biography	Wikified topics
Big fan of l.a.lakers. Love hunting and fishing	*Los Angeles Lakers*, *Hunting*, *Fishing*
half japanese, like piano, violin	*Piano*, *Violin*

Table 2: Some biographies and extracted interests.

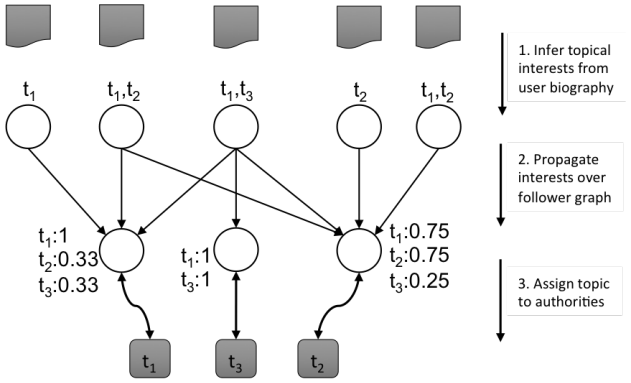

Figure 3: High level overview of ALF.

then used a named entity detection model (see, for example, [15, 17, 9]) to identify entities (interests) mentioned in the biographies of users and selected those that belonged to at least one of the whitelisted categories. This yielded high-precision interests $I(\cdot)$ for many users. Table 2 lists some examples of inferred interests from the biographies. Finally, we set $\mathcal{T} = \bigcup_{u \in V} I(u)$.

4.2 Interest Propagation over Follower Graph

From the known interests of a few users, we must estimate authority scores for all users. The standard algorithm in such cases is label propagation [39, 40], which works as follows. Consider a $|\mathcal{T}| \times |V|$ real valued matrix $\boldsymbol{S^c}$ where $\boldsymbol{S^c_{tu}}$ is clamped to 1 if user u is interested in topic t, i.e., $t \in I(u)$, otherwise it is left empty. The goal is to build a matrix $\boldsymbol{S}$ so as to minimize $C(\boldsymbol{S}) = \sum_{(u \rightarrow v) \in E} ||\boldsymbol{S}_u - \boldsymbol{S}_v||^2$, while ensuring that the known interests $\boldsymbol{S^c}$ are retained in $\boldsymbol{S}$; here, $\boldsymbol{S}_u$ is the column vector of $\boldsymbol{S}$ and $||v||$ is the 2-norm of vector v. $C(\boldsymbol{S})$ can be minimized by solving the fixed point equations $\boldsymbol{S}_v = \frac{1}{n_v^{in} + n_v^{out}} \left[\sum_{u \rightarrow v} \boldsymbol{S}_u + \sum_{v \rightarrow w} \boldsymbol{S}_w\right]$. However, this is ill-suited to our problem: (a) authority scores are considered identical to topical interest scores, which is not true, and (b) this approach can be computationally intensive given the scale of Instagram, as it might require many map-reduce rounds over the follower graph until convergence.

Even if we created a separate matrix $\boldsymbol{F}$ of authority scores, and tried to infer both $\boldsymbol{S}$ and $\boldsymbol{F}$ by minimizing the function $\sum_{(u \rightarrow v) \in E} ||\boldsymbol{S}_u - \boldsymbol{F}_v||^2$, this runs into two problems. First, setting both $\boldsymbol{S}$ and $\boldsymbol{F}$ to the all-ones matrix is a solution. Even if the objective is regularized to prevent this, the results are not easily *explainable*: the authority scores of node v can depend heavily on the interests of nodes far from the local neighborhood of v. However, simply restricting propagation to the local neighborhood risks losing the power and advantages of label propagation. Instead, we propose a method to find *explainable* and *broader* inferences, that can then be weighted depending on the business needs.

Specifically, we split interests $\boldsymbol{S}$ into the known interests $\boldsymbol{S^c}$ and the "*broader*" interests $\boldsymbol{S^i}$. Similarly, the authority scores $\boldsymbol{F}$ are split into "explainable" scores $\boldsymbol{F^e}$ and "broader" scores $\boldsymbol{F^i}$. The explainable authority scores $\boldsymbol{F^e}$ must be based only on known interests $\boldsymbol{S^c}$, while the broader interests $\boldsymbol{S^i}$ and scores $\boldsymbol{F^i}$ must be consistent with each other. Finally, we link the broader and explainable terms by requiring $\boldsymbol{S^i}$ to be close to that expected from $\boldsymbol{F^e}$. This leads

to the following objective:

$$\text{Minimize } \frac{1}{2} \sum_{(u\to v)\in E} \Big[||\boldsymbol{F_v^e} - \boldsymbol{S^c}_u||^2 + \alpha \cdot ||\boldsymbol{F_v^e} - \boldsymbol{S_u^i}||^2 + \beta \cdot ||\boldsymbol{F_v^i} - \boldsymbol{S_u^i}||^2 \Big] \quad (3)$$

The parameters α and β trade-off the importance of matching the explainable terms and the inferred terms. Finally, the authority score are a combination of the explainable and inferred scores $\boldsymbol{F} = \boldsymbol{F^e} + \gamma \cdot \boldsymbol{F^i}$, where the parameter γ is chosen based on business concerns, such as the required degree of explainability of results.

Let $\boldsymbol{A}$ be the adjacency matrix of graph $\mathcal{G}$.

THEOREM 1. *Under the objective of Eq. 3, we have*

$$\boldsymbol{F} = \frac{1}{1+\alpha}\boldsymbol{S^c}\left[I + \frac{\gamma+\alpha(1+\gamma)}{(1+\alpha)(1+\beta/\alpha)}\boldsymbol{M}(I-\kappa\boldsymbol{M})^{-1}\right]\boldsymbol{P}_{\rightarrow}$$

$$\boldsymbol{S^i} = \frac{1}{(1+\alpha)(1+\beta/\alpha)}\boldsymbol{S^c}\boldsymbol{M}\left[I-\kappa\boldsymbol{M}\right]^{-1}$$

where

$$\boldsymbol{D}_{in} = diag(\mathbf{1}^t\boldsymbol{A}) \qquad \boldsymbol{P}_{\rightarrow} = \boldsymbol{A}\boldsymbol{D}_{in}^{-1}$$

$$\boldsymbol{D}_{out} = diag(\boldsymbol{A}\mathbf{1}) \qquad \boldsymbol{P}_{\leftarrow} = \boldsymbol{A}^t\boldsymbol{D}_{out}^{-1}$$

$$\kappa = \left(\frac{\alpha}{1+\alpha}+\frac{\beta}{\alpha}\right)\left(1+\frac{\beta}{\alpha}\right)^{-1} \qquad \boldsymbol{M} = \boldsymbol{P}_{\rightarrow}\boldsymbol{P}_{\leftarrow},$$

The operator $\boldsymbol{P}_{\rightarrow}$ corresponds to propagating labels from $\boldsymbol{S}$ to $\boldsymbol{F}$, while $\boldsymbol{P}_{\leftarrow}$ corresponds to the opposite propagation direction. $\boldsymbol{M}$ corresponds to a combination of the forward and backward pass.

Since matrix inversion becomes difficult for large matrices, a multi-pass solution is suggested by the following corollary.

COROLLARY 1. *When $0 < \beta \ll 1$, $0 < \alpha \ll \min\{1, \gamma\}$,*

$$\boldsymbol{F} \approx \boldsymbol{S^c}\boldsymbol{P}_{\rightarrow} + \gamma\frac{\alpha}{\beta}\boldsymbol{S^c}\left[\sum_j \left(\frac{\beta}{\alpha+\beta}\boldsymbol{M}\right)^j\right]\boldsymbol{P}_{\rightarrow}.$$

Thus, the general solution can be found by a *weighted* label propagation where a factor of $\sqrt{\beta/(\alpha+\beta)}$ is used to dampen successive iterations. This prevents the interests of far-off nodes from affecting authority scores too much, and keeps it grounded in the interests of nodes in the local neighborhood.

Practical Instantiation of the Propagation Algorithm

Running multiple passes of the propagation algorithm can be computationally intensive in large networks. We propose Algorithm 1 which works well in practice with just 3 passes. In the first pass, it computes the fraction of followers of v interested in topic t w.r.t. the followers who express some interests. In the second pass, interests of all users are re-estimated, thereby increasing the coverage to all users. Finally, the algorithm computes the label scores from the inferred interests of all the followers. The algorithm only requires 3 passes over the follower graph and in this sense it is quite efficient. We also note that it is easy to parallelize and it scales well to handle massive datasets.

The following corollary establishes the connection between the solution of Eq. 3 and Algorithm 1.

Algorithm 1 Fast Algorithm for Interest Propagation

Set $\boldsymbol{F^e} = 0$ and $\boldsymbol{F^i} = 0$.

PASS 1:
Define $C(\boldsymbol{F^e}) = \frac{1}{2}\sum_{\substack{(u\to v)\in E \\ I(u)\neq\phi}} ||\boldsymbol{F_v^e} - \boldsymbol{S^c}_u||^2$. The minimizer of $C(\boldsymbol{F^e})$ can be computed in closed form:

$$\boldsymbol{F_v^e} = \frac{1}{m_v^{in}} \sum_{\substack{(u\to v)\in E \\ I(u)\neq\phi}} \boldsymbol{S^c}_u, \quad (4)$$

where $m_v^{in} = |\{u : (u \to v) \in E \wedge I(u) \neq \phi\}|$

PASS 2:
Define $C(\boldsymbol{S^i}) = \frac{1}{2}\sum_{(u\to v)\in E} ||\boldsymbol{F_v^e} - \boldsymbol{S_u^i}||^2$. Compute minimizer of $C(\boldsymbol{S^i})$ as $\boldsymbol{S_u^i} = \frac{1}{n_u^{out}}\sum_{(u\to v)\in E}\boldsymbol{F_v^e}$.

PASS 3:
Define $C(\boldsymbol{F^i}) = \frac{1}{2}\sum_{(u\to v)\in E} ||\boldsymbol{F_v^i} - \boldsymbol{S_u^i}||^2$. Compute minimizer of $C(\boldsymbol{F^i})$ as $\boldsymbol{F_v^i} = \frac{1}{n_v^{in}}\sum_{(u\to v)\in E}\boldsymbol{S_u^i}$.

Return $\boldsymbol{F} = \boldsymbol{F^e} + \boldsymbol{F^i}$.

COROLLARY 2. *When $\beta \ll \alpha \ll 1$ and $\gamma = 1$, we have $\boldsymbol{F} \approx \boldsymbol{S^c}[I + \boldsymbol{M}]\boldsymbol{P}_{\rightarrow}$, which is the same result as Algorithm 1.*

Thus, Algorithm 1 solves the setting where explainability of $\boldsymbol{F^e}$ and $\boldsymbol{S^i}$ in terms of the clamped interests is particularly valued, and the final authority score $\boldsymbol{F}$ weighs the explainable part $\boldsymbol{F^e}$ and the inferred part $\boldsymbol{F^i}$ equally.

4.3 Estimating Topical Authorities

There are three steps in estimating authority scores and assigning authority topics to users. These steps are described below.

4.3.1 Normalized Label Scores

Algorithm 1 ensures that $\boldsymbol{F}_{tu}$ is high if u is a known authority on t, in keeping with Hypothesis 1. However, it provides no guarantees about the scores for topics where u is not an authority. In fact, we notice that popular topics have high $\boldsymbol{F}$ scores in general, since most people are interested in those topics. Hence, a naive authority selection method that assigns topic $\arg\max_t\{\boldsymbol{F}_{tu}\}$ to user u would end up saying most users are authorities on popular topics. To address this issue, we must normalize the authority scores per topic relative to other users.

In general, we would proceed by computing the cumulative density, as follows:

$$P_F(u|t) = \frac{1}{|V|}\sum_{v\in V}\mathbf{1}[\boldsymbol{F}_{tu} > \boldsymbol{F}_{tv}], \quad (5)$$

where $\mathbf{1}[cond]$ is the indicator random variable which is 1 if *cond* is true, otherwise 0. P_F defines relative standing of users per topic. However computing the *cdf* function takes $\mathcal{O}(|\mathcal{T}|\cdot|V|\cdot\log|V|)$ time, which is computationally intensive. However, we make the following observation.

OBSERVATION 2. *The rows of $\boldsymbol{F}$ are log-normally distributed.*

Figure 4 confirms this trend for the basketball topic. This observation simplifies our computations considerably. We

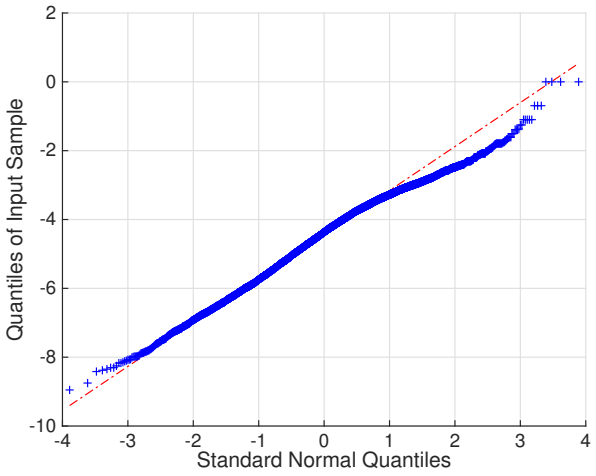

Figure 4: Quantile plot of $Log(F)$ for topic *basketball*. We randomly picked $10,000$ users for this plot.

compute the sufficient statistics per topic,

$$\boldsymbol{\mu} = \frac{\boldsymbol{L}\mathbf{1}}{|V|}, \qquad \boldsymbol{\sigma} = \sqrt{\frac{diag([\boldsymbol{L} - \boldsymbol{\mu}\mathbf{1}^t][\boldsymbol{L} - \boldsymbol{\mu}\mathbf{1}^t]^t)}{|V|}}$$

where $\boldsymbol{L} = \log\{\boldsymbol{F}\}$. The sufficient statistics $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ can be computed efficiently in $|\mathcal{T}| \cdot |V|$ time. The relative topic scores are then computed for user u through the z-score normalization scheme:

$$\boldsymbol{ZF}_u = diag(\boldsymbol{\sigma})^{-1}(\boldsymbol{L}_u - \boldsymbol{\mu}). \tag{6}$$

$\boldsymbol{ZF}$ represents the relative topical authority score of users.

4.3.2 Computing Authority Score

The z-scoring technique provides a way to compare how a user fares on different topics. However, for users with a modest number of followers, it biases the computation towards tail (less popular) topics. We illustrate this problem via the three topics in Table 3. The topics have very different popularities but nearly identical $\boldsymbol{\sigma}$. However, mean $\boldsymbol{\mu}$ increases as popularity increases, which can propel a tail topic's z-score over a that of a popular topic. For instance, consider a tail topic t_{tail} and a popular topic t_{pop} with $\sigma_{tail} \sim \sigma_{pop}$ and $\mu_{tail} = \mu_{pop} - 4$. For an expert u on t_{pop} to be labeled accurately by z-score, we must have:

$$\frac{\boldsymbol{F}_{t_{pop}u}}{\boldsymbol{F}_{t_{tail}u}} > 10^{\mu_{t_{pop}} - \mu_{t_{tail}}} = 10,000. \tag{7}$$

For users with even 10^4 followers, clearing the above threshold is not possible, unless $\boldsymbol{F}_{t_{tail}u} = 0$. Clearly, for a moderately popular account, satisfying the above inequality is a tall order. Our solution is to weight $\boldsymbol{ZF}$ with the number of topical followers, as follows:

$$\boldsymbol{wZF}_u = diag(\boldsymbol{ZF}_u \log(n_u^{in}\boldsymbol{F}_u)^t). \tag{8}$$

This weighted z-score $\boldsymbol{wZF}$ solves the problem of a low popularity topic bumping up without merit. A second benefit is that it provides an intuitive ordering of top-ranked users for each topic; ordering based on $\boldsymbol{ZF}$ alone is not useful for recommendation as it is susceptible to placing low popularity accounts over popular ones. This issue is elegantly addressed by the $\boldsymbol{wZF}$ scheme which combines the z-score with the topical popularity of the account – providing a robust ordering.

Topic name	Popularity	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$
Music	High	-3.04	0.51
Comedy	Medium	-5.84	0.52
Planet	Low	-7.62	0.55

Table 3: Statistics of some topics.

4.3.3 Eliminating False Positives

We use $\boldsymbol{wZF}$ to assign topics to users (u is assigned topic $\arg\max_t\{\boldsymbol{wZF}_{tu}\}$). However, downstream applications may require *high precision*; for example, a recommendation system based on authority detection would require high confidence in authority assignments. Hence, we need a post-processing step to filter out false positives. Although, $\boldsymbol{wZF}$ mitigates the false positive issue to a large extent, it does not resolve it completely. Here we identify the two main types of false positives that are not yet addressed.

FP1 *Tail user with low authority scores.* Users with moderate to low follower count can crowd a popular topic.

FP2 *Celebrities with high authority scores.* Certain celebrities that are followed by users with different cross-sections of interests can be assigned wrong topics.

Solving FP1. $FP1$ is characterized with low $\boldsymbol{wZF}$ scores which can be effectively addressed by filtering assignments that fall below a certain threshold. A standard way of computing the threshold is by picking scores that are above a fixed percentile level ρ. Let the sorted set of authorities for topic t be the users $m_1, m_2, \ldots, m_{n_t}$, where n_t is the number of users assigned as authorities on t and $\boldsymbol{wZF}_{t,m_i} \geq \boldsymbol{wZF}_{t,m_j}$ for $i < j$. The threshold θ_t for topic t is then defined as

$$\theta_t = \boldsymbol{wZF}_{t,q_t}$$
$$\text{where } q_t = \left\lfloor \frac{\rho}{100} n_t \right\rfloor.$$

While this is intuitively appealing, no single percentile level ρ works well over the entire range of topics. This is because for topics with very large n_t, θ_t values turn out to be very low – amounting to an ineffective filtering for these topics. If ρ is decreased to take care of this issue, then it would result in a very aggressive filtering for topics with low n_t. Instead, we divide the topics into three buckets: *popular*, *mid*, and *tail* based on their n_t values, and use a separate percentile level per bucket. For example, for a popular topic, the percentile level ρ_{pop} would be used for filtering. The percentile level for the buckets follows the constraint:

$$\tau\rho_{pop} = \rho_{mid} = \rho_{tail}/\tau, \tag{9}$$

where we set $\rho_{mid} = 60$ and $\tau = 1.5$.

Solving FP2. Celebrity false positives are characterized with high $\boldsymbol{wZF}$ scores. Hence the thresholding that is applied for filtering FP1 does not work for this case. Instead, we consider a voting between the different scores obtained thus far. Since $\boldsymbol{wZF}$ is already used for assigning authority topics to users, we consider $\boldsymbol{F}$ and $\boldsymbol{ZF}$. For assignment

Algorithm 2 Authority Based Recommender

Require: $\mathcal{A}, u, \hat{a}, \hat{b}$
  $\Phi = \{\}$
  for $t \in \mathcal{T}$ **do**
    $\Phi_t = \{\}$
    if $a_{tu} > \hat{a}$ **then**
      for $v \in \mathcal{A}_t$ and $|\Phi_t| < \hat{b}$ **do**
        if $u \not\rightarrow v$ **then**
          $\Phi_t = \Phi_t \cup \{v\}$
        end if
      end for
    end if
  end for
  return $\bigcup_{t \in \mathcal{T}} \Phi_t$

(x, t) obtained from $\boldsymbol{wZF}$, if t appears within top k of both the scorers $\boldsymbol{F}$ and $\boldsymbol{ZF}$, the authority assignment is retained, otherwise it is discarded.

Intuitively, if the celebrity is assigned a niche topic, then we expect that topic to not appear in the top-k of her $\boldsymbol{F}$ score. On the other hand, if she is assigned a popular topic, then we have similar expectation from the $\boldsymbol{ZF}$ score. Empirically, we find that $k = 5$ works best.

Time Complexity of ALF

The time complexity of ALF is $\mathcal{O}(|\mathcal{T}| \cdot (|V| + |E|))$. This is because our interest propagation algorithm runs in $\mathcal{O}(|\mathcal{T}| \cdot |E|)$ time. From computation of authority scores to elimination of false positives the time complexity is $\mathcal{O}(|\mathcal{T}| \cdot |V|)$. We put ALF to practice for the large scale at Instagram through Apache Hive[3].

5. AUTHORITY BASED RECOMMENDER

The output of ALF model is an ordered authority list ($\mathcal{A}$) for each topic t in $\mathcal{T}$, ordered in decreasing order of $\boldsymbol{wZF}_{t\cdot}$ scores for users that are assigned that topic. Now, a user's enthusiasm for topic t can be judged by the number of authorities on t that she follows.

$$\text{Enthusiasm } a_{tu} = |\{v : v \in \mathcal{A}_t \wedge u \rightarrow v\}| \qquad (10)$$

If a_{tu} is greater than a specified threshold $\hat{a}$ then we consider u to be highly enthusiastic about t. In this case, the top $\hat{b}$ relevant authorities in t that u is not already following are recommended to her. Algorithm 2 details this process.

6. EXPERIMENTAL EVALUATION

We provide four different evaluations to test the effectiveness of our model. First, we report our performance compared to other state of the art baseline models in a "*what users to follow*" suggestion task in an actual production environment. Second, we provide a controlled comparison of the best performing baseline and ALF in a live experimental setting. Third, we compare ALF to several benchmark models in a recall task that utilizes an *expert curated* list of topical authorities. Finally, we report a manual validation of the top accounts identified by our approach across 120 different topics and 24,000 labeled accounts.

[3]`https://hive.apache.org/`

Model	CTR	Conversion
ALF	**1.0**	**1.0**
NN-based	0.68	0.71
MF-based	0.45	0.41
Hybrid	0.79	0.88
Graph-based	0.82	0.75

Table 4: Performance of the best performing recommendation model within each category for the user recommendation task in the production environment. For a relative comparison, the performance numbers are normalized w.r.t. ALF.

6.1 User Recommendation Task

Our first goal is to test the performance of our model for the task of recommending users to follow. We compare our model against several fine-tuned baseline models in an actual production environment. We present a high-level categorization of the baseline models.

- *NN-based*: This category comprises of nearest neighbor (NN) based collaborative-filtering (CF) models to compute user similarity[4] (e.g. [18, 35]). For a general survey on other recommendation methods, see [27, 2].
- *MF-based*: This family of models uses matrix factorization based methods for recommendation (see for example PMF [34], Koren et al. [26]).
- *Hybrid*: These models combine content based methods with collaborative filtering methods for recommendation (see for example [29, 26, 24]).
- *Graph-based*: These models recommend using graph based features such as PageRank [8], preferential attachment [4], node centrality, friends of friends, etc.

First, each model generates k recommendations per user in realtime. The generated recommendations from all the models are then mixed together and an independent ranker orders them. Finally the ordered recommendations are shown to the end user. We measure the performance of a recommender on two criteria: (1) *click through rate* (CTR), which is the observed probability of users clicking the recommendations, and (2) *conversion rate*, which is the observed probability of users actually choosing to follow the recommended account. For a fair evaluation, we account for the position bias effect [21] by measuring the performance of a model only if one of its recommendations is shown in the top position. The recommendations are shown over a 1-week period to all Instagram users.

Table 4 shows the relative performance of different models in comparison to ALF. We observe that ALF performs better than all the baseline models. The performance numbers are significant using one-sided t-test with $p = 0.001$. The result shows that in a live production setting, our model is able to generate more useful recommendations in comparison to all the fine-tuned baseline methods.

6.2 Recommendation in a Controlled Setting

The previous experiment measured the performance of the models in-the-wild, i.e., the recommendations from all the

[4]Similarity can be computed on the basis of co-likes, co-follows, co-occurrence of hashtags or interests.

Model	CTR	Conversion	Participation
ALF	*1*	*1*	*1*
Hybrid	0.84	0.89	0.95

Table 5: Performance of the best baseline in comparison to ALF in a controlled production environment.

models were competing against one another simultaneously. Next, we consider a controlled setting in which we compared our model with the best baseline model (*Hybrid*) in a randomized trial. The randomized trial overcomes the confounding bias and helps in the attribution of user participation[5] increase directly to the underlying model.

We perform A/B testing using a block randomized trial on a 5% random sample of Instagram users. The users are split into treatment and control groups, while controlling for the population distribution within the two groups. The control group is shown the recommendations generated by the best baseline (*Hybrid*) while the treatment group is shown recommendations by ALF. Apart from *CTR* and *conversion*, we also measure the increase in user participation once they acted on the recommendations. We ran this experiment for a 1-week period.

Table 5 shows that our model performs better than the best baseline model (with $p = 0.001$ using one-sided t-test). In particular, the improvement in user participation indicates that indeed ALF generates recommendations that are more appealing to the end-users[6].

6.3 Precision and Recall Comparison

Here we compare our model with prior state-of-art models over a labeled dataset. This curated set consists of 25 topics, with 15 must-follow authorities on each of those topics. We use this dataset to perform a detailed comparison against a broader class of models, and also to test variants of ALF. The models we tested were the following:

- *TwitterRank*: We constructed the topically weighted follower graph based on the similarity of topical activity of two nodes and used the TwitterRank [37] algorithm over the topical graph to identify the topical authorities.
- *Hashtags*: This baseline uses the approach proposed by Pal et al. [31]. Here we consider the hashtags from the content generated by the users and generate several graph-based and nodal metrics for the users and ran the proposed ranker.
- *LDA*: Each user is associated with a "document" containing the biographies of all her followers. LDA is run on these documents, and LDA topics that closely match the 15 labeled topics are manually identified. Next, each user is associated with several features, including the LDA topic probabilities, number of followers for each topic, and features obtained from Hashtags and the follower graph. The relative importances of these features are learnt by multinomial logistic regression [5] using 5 topics and their 15 known authorities as positive examples. These features are then used to rank the authorities for the remaining 20 labeled topics. Experiments were repeated with different train/test splits on topics.
- *PageRank*: This baseline uses PageRank [8] over the follower graph. For each topic t, a separate iteration of the PageRank algorithm is run after initializing the PageRank of user u to 1 if u mentions topic t in her biography (i.e. $\boldsymbol{S^c_{tu}} = 1$). Finally, a user is assigned the topic for which she has the highest PageRank.
- *Likes only*: This method extracts users' interests based on the content liked by them and then runs ALF on these interests.
- *Posts only*: This method extracts users' interests based on the content generated by them and then runs ALF on these interests.
- *No Wiki*: This method considers all the unigrams from the users' biography as interests and then runs ALF on these interests.
- *No weighting*: This baseline is based on ALF with a difference that users were scored based on their $\boldsymbol{ZF}$ score instead of $\boldsymbol{wZF}$.

Method	Precision	Recall	F1 = $\frac{2PR}{P+R}$
TwitterRank	0.81	0.31	0.45
Hashtags	0.84	0.26	0.40
LDA	0.68	0.18	0.28
PageRank	0.85	0.21	0.34
Likes only	0.96	0.54	0.69
Posts only	0.92	0.56	0.70
No Wiki	0.91	0.41	0.56
No weighting	0.96	0.57	0.72
ALF	**0.96**	**0.66**	**0.78**

Table 6: Precision and Recall of different models over the label dataset. We set $k = 200$ to compute the performance of the models.

Performance Metric: We compare the performance of the models on the basis of their precision and recall. Let t denote a topic from the label dataset and $\mathcal{B}_t$ denote the set of authorities on t as identified in the curated dataset. Let $\mathcal{A}^k_t$ denote the top k authorities on t discovered by a given model. Model precision and recall is then defined as follows:

$$\text{Precision}_k = \frac{|\mathcal{A}^k_t \cap \mathcal{B}_t|}{\left|\mathcal{A}^k_t \cap \left(\bigcup_t \mathcal{B}_t\right)\right|} \qquad \text{Recall}_k = \frac{|\mathcal{A}^k_t \cap \mathcal{B}_t|}{|\mathcal{B}_t|}$$

We note that we must use a non-standard measure of precision since the curated list of authorities is not comprehensive, so a model's precision should only be measured over the authorities that are labeled.

We pick top $k = 200$ authorities per model. Table 6 shows the performance of the different models. The result shows that our model has the highest precision. This is intuitively expected as we take steps to ensure that false positives are eliminated. However it also has the highest recall, which shows its effectiveness at discovering topical authorities.

[5]Number of likes and comments within a login session account for the participation.

[6]We note in passing that the numbers in Tables 4 and 5 are not directly comparable due to the confounding effects of other methods in Table 4.

In terms of the performance of variants of ALF, we notice that all of them have high precision. However the recall varies. The models based on the users' production or consumption data have much lower recall, confirming our initial assessment that models based on users' activity might not work as well for this domain. We also note that the PageRank based model does not work as well due to the concentration of scores at nodes with large in-degree. We also note that z-scoring without weighting by follower counts has lower recall than ALF. Overall, the results emphasize the fact that users' biographies are a more effective estimator of their interests than their activity.

6.4 Qualitative Model Performance

The experiments so far establish the effectiveness of ALF for the recommendation task and in surfacing well-known topical authorities. Here we estimate the qualitative performance of the model using domain experts. For this, we selected the most popular 120 topics discovered by ALF and top 200 authorities identified by ALF per topic. The popularity of a topic is defined based on the number of users enthusiastic about that topic (see Eq. 10). The resulting dataset consists of $24,000$ authorities.

The expert evaluators were asked to evaluate based on the public content of the authorities "whether a user is an authority on the assigned topic or not". The expert assessment yielded a 94% accuracy score for ALF. The high accuracy level over this large labeled dataset is consistent with the precision of ALF over the labeled dataset. This result highlights the efficacy of ALF for authority discovery in Instagram.

7. CONCLUSIONS

In this paper, we presented an Authority Learning Framework (ALF) which is based on the self-described interests of the followers of popular users. We proposed a generalized label propagation algorithm to propagate these interests over the follower graph and proposed a practical instantiation of it that is practically feasible and effective. We also showed how authority scores can be computed from the topic specific normalization and how different types of false positives can be eliminated to obtain high quality topic authority lists.

We conducted rigorous experiments in production setting and over a hand-curated dataset to show the effectiveness of ALF over competitive baseline methods for the user recommendation task. Qualitative evaluation of ALF showed that it yields high precision authority lists. As part of future work, we would like to combine variants of ALF and examine its performance for the user recommendation task.

8. APPENDIX

Proof of Theorem 1. By setting to zero the derivatives of the objective (Eq. 3) with respect to $\boldsymbol{F^i}$, $\boldsymbol{S^i}$, and $\boldsymbol{F^e}$ respectively, we find:

$$\boldsymbol{F_v^i} = \frac{\sum_{\{u|u\to v\in E\}} \boldsymbol{S_u^i}}{n_v^{in}} \tag{11}$$

$$\boldsymbol{S_u^i} = \frac{\sum_{\{v|u\to v\in E\}} \left[\boldsymbol{F_v^e} + \beta/\alpha \boldsymbol{F_v^i}\right]}{n_u^{out}\cdot(1+\beta/\alpha)} \tag{12}$$

$$\boldsymbol{F_v^e} = \frac{\sum_{\{u|u\to v\in E\}} \left[\boldsymbol{S_u^c} + \alpha\cdot \boldsymbol{S_u^i}\right]}{n_v^{in}\cdot(1+\alpha)} \tag{13}$$

These may be written in matrix form as follows:

$$\boldsymbol{F^i} = \boldsymbol{S^i A D_{in}^{-1}} \tag{14}$$

$$\boldsymbol{S^i} = \frac{1}{1+\beta/\alpha}\left(\boldsymbol{F^e} + \frac{\beta}{\alpha}\boldsymbol{F^i}\right)\boldsymbol{A^t D_{out}^{-1}} \tag{15}$$

$$\boldsymbol{F^e} = \frac{1}{1+\alpha}\left(\boldsymbol{S^c} + \alpha \boldsymbol{S^i}\right)\boldsymbol{A D_{in}^{-1}} \tag{16}$$

Substituting into equation 15, we find:

$$(1+\beta/\alpha)\,\boldsymbol{S^i} = \left[\left(\frac{\boldsymbol{S^c}+\alpha\boldsymbol{S^i}}{1+\alpha}\right)\boldsymbol{P_\to} + \frac{\beta}{\alpha}\boldsymbol{S^i P_\to}\right]\boldsymbol{P_\leftarrow}$$

$$\Rightarrow \boldsymbol{S^i} = \frac{1}{(1+\alpha)(1+\beta/\alpha)}\boldsymbol{S^c P_\to P_\leftarrow} + \kappa \boldsymbol{S^i P_\to P_\leftarrow}$$

$$\Rightarrow \boldsymbol{S^i} = \frac{1}{(1+\alpha)(1+\beta/\alpha)}\boldsymbol{S^c M}\left[I-\kappa\boldsymbol{M}\right]^{-1}$$

Substituting back into Eqs. 14 and 16, we get the equations for $\boldsymbol{F^e}$ and $\boldsymbol{F^i}$. Now, using $\boldsymbol{F} = \boldsymbol{F^e} + \gamma\boldsymbol{F^i}$ yields the desired result.

To show that the inverse always exists, note that $0 \le \kappa < 1$. Also, the entries of $\boldsymbol{M}$ are given by

$$\boldsymbol{M}_{ij} = \sum_k \frac{\boldsymbol{A}_{ik}\boldsymbol{A}_{jk}}{n_k^{in} n_j^{out}}.$$

Hence, the row-sum of row i in $\boldsymbol{M}$ is $\sum_j \boldsymbol{M}_{ij} = 1$, which is identical for every row. Hence, by the Perron-Frobenius Theorem, the maximum eigenvalue of $\boldsymbol{M}$ is 1. Hence, the maximum eigenvalue of $\kappa\boldsymbol{M}$ is $\kappa < 1$. Hence, the inverse of $I - \kappa\boldsymbol{M}$ exists. □

Proof of Corollary 1. Applying $\alpha \ll \min\{1,\gamma\}$ and $\beta \ll 1$ to Theorem 1, we find:

$$\boldsymbol{F} \approx \boldsymbol{S^c}\left[I + \frac{\gamma}{1+\beta/\alpha}\boldsymbol{M}\left(I-\kappa\boldsymbol{M}\right)^{-1}\right]\boldsymbol{P_\to} \tag{17}$$

$$= \boldsymbol{S^c P_\to} + \frac{\gamma}{1+\beta/\alpha}\boldsymbol{M}\left(I-\kappa\boldsymbol{M}\right)^{-1}\boldsymbol{P_\to} \tag{18}$$

Under the conditions of the Corollary, we observe that $\kappa \approx \beta/(\alpha+\beta)$. Now, doing a Neumann series expansion of the inverse, we get:

$$\boldsymbol{F} \approx \boldsymbol{S^c P_\to} +$$

$$\frac{\gamma}{1+\beta/\alpha}\left(\boldsymbol{M} + \frac{\beta}{\alpha+\beta}\boldsymbol{M}^2 + \left(\frac{\beta}{\alpha+\beta}\right)^2\boldsymbol{M}^3 + \ldots\right)\boldsymbol{P_\to}$$

$$= \boldsymbol{S^c P_\to} + \gamma\frac{\alpha}{\beta}\boldsymbol{S^c}\left[\sum_j\left(\frac{\beta}{\alpha+\beta}\boldsymbol{M}\right)^j\right]\boldsymbol{P_\to} \tag{19}$$

□

Proof of Corollary 2. Under the conditions of the Corollary, $\beta/\alpha \approx 0$ and hence $\kappa \approx 0$. Hence, from Thm. 1, we find:

$$\boldsymbol{F} \approx \boldsymbol{S^c}\left[I+\boldsymbol{M}\right]\boldsymbol{P_\to}.$$

Now, consider Algorithm 1. Pass 1 corresponds to the calculation of $\boldsymbol{F^e} = \boldsymbol{S^c P_\to}$. Pass 2 sets $\boldsymbol{S^i} = \boldsymbol{F^e P_\leftarrow} = \boldsymbol{S^c M}$ (since $\boldsymbol{M} = \boldsymbol{P_\to P_\leftarrow}$). Finally, pass 3 sets $\boldsymbol{F^i} = \boldsymbol{S^i P_\to} = \boldsymbol{S^c M P_\to}$. Hence, the final computation of $\boldsymbol{F}$ yields $\boldsymbol{F} = \boldsymbol{F^e} + \boldsymbol{F^i} = \boldsymbol{S^c}\left[I+\boldsymbol{M}\right]\boldsymbol{P_\to}$, as desired. □

9. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM*, 2008.

[2] X. Amatriain, A. Jaimes, N. Oliver, and J. Pujol. Data mining methods for recommender systems. In *Recommender Systems Handbook*. Springer, 2011.

[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR*, 2006.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 1999.

[5] A. L. Berger, S. D. Pietra, and V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1996.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[7] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM TOIT*, 2005.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 1998.

[9] Z. Cai, K. Zhao, K. Q. Zhu, and H. Wang. Wikification via link co-occurrence. In *CIKM*, 2013.

[10] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM*, 2003.

[11] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 1990.

[12] Y. Ding and J. Jiang. Extracting interest tags from twitter user biographies. In *Information Retrieval Technology*. 2014.

[13] A. Farahat, G. Nunberg, and F. Chen. Augeas: authoritativeness grading, estimation, and sorting. In *CIKM*, 2002.

[14] D. Fisher, M. Smith, and H. T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. *HICSS*, 2006.

[15] A. Gattani, D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *PVLDB*, 2013.

[16] S. Ghosh, N. K. Sharma, F. Benevenuto, N. Ganguly, and P. K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*, 2012.

[17] Z. Guo and D. Barbosa. Robust entity linking via random walks. In *CIKM*, 2014.

[18] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *SIGIR*, 1999.

[19] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[20] A. Java, P. Kolari, T. Finin, and T. Oates. Modeling the spread of influence on the blogosphere.

[21] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM TOIS*, 2007.

[22] P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In *CIKM*, 2007.

[23] D. Kempe. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[24] B. M. Kim, Q. Li, C. S. Park, S. G. Kim, and J. Y. Kim. A new approach for combining content-based and collaborative filters. *J. Intell. Inf. Syst.*, 2006.

[25] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA*, 1998.

[26] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*, 2008.

[27] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 2009.

[28] X. Liu, W. B. Croft, and M. B. Koll. Finding experts in community-based question-answering services. In *CIKM*, 2005.

[29] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *National Conference on Artificial Intelligence*, 2002.

[30] A. Pal. Discovering experts across multiple domains. In *SIGIR*, 2015.

[31] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM*, 2011.

[32] A. Pal, F. M. Harper, and J. A. Konstan. Exploring question selection bias to identify experts and potential experts in community question answering. *TOIS*, 2012.

[33] A. Popescu, K. Y. Kamath, and J. Caverlee. Mining potential domain expertise in pinterest. In *Workshop Proceedings of UMAP*, 2013.

[34] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, 2007.

[35] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, 2001.

[36] X. Tang, M. Zhang, and C. C. Yang. User interest and topic detection for personalized recommendation. In *Web Intelligence*, 2012.

[37] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.

[38] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW*, 2007.

[39] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. *Technical Report, Carnegie Mellon University*, 2002.

[40] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.