Social Factors in Closed-Network Content Consumption

Parisa Kaghazgaran*
Texas A&M University
College Station, TX
kaghazgaran@tamu.edu

Maarten Bos, Leonardo Neves and Neil Shah
Snap Inc.
Santa Monica, CA
{maarten, lneves, nshah}@snap.com

ABSTRACT

How do users on social platforms consume content shared by their friends? Is this consumption socially motivated, and can we predict it? Considerable prior work has focused on inferring and learning user preferences with respect to broadcasted, or open-network content in public spheres like webpages or public videos. However, user engagement with narrowcasted, closed-network content shared by their friends is considerably under-explored, despite being a commonplace activity. Here we bridge this gap by focusing on consumption of visual media content in closed-network settings, using data from Snapchat, a large multimedia-driven social sharing service with over 200M daily active users. Broadly, we answer questions around content consumption patterns, social factors that are associated with such consumption habits, and predictability of consumption time. We propose models for patterns in users' time-spending behaviors across friends, and observe that viewers preferentially and consistently spend more time on content from certain friends, even without considering any explicit notion of intrinsic content value. We also find that consumption time is highly correlated with several engagement-based social factors, suggesting a large social role in closed-network content consumption. Finally, we propose a novel approach of modeling future consumption time as a learning-to-rank task over users' friends. Our results demonstrate significant predictive value (0.815 P@1, 0.650 nDCG@10) using only social factors. We expect our work to motivate additional research in modeling consumption and ranking of online closed-network content.

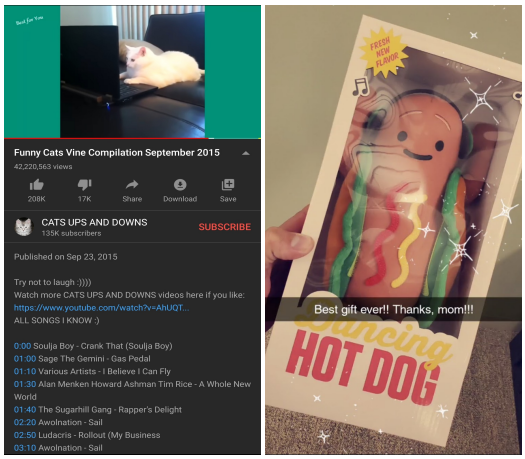
CCS CONCEPTS

• **Information systems → Collaborative and social computing systems and tools**; **Content ranking**; **Social networks**; *Personalization*.

ACM Reference Format:
. 2020. Social Factors in Closed-Network Content Consumption. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411935

1 INTRODUCTION

Social platforms offer a multitude of ways for users to interact and communicate with each other. Interactions vary from mechanisms as light as likes, pokes, and upvotes to ones as involved as text messaging, media sharing, and monetary transfers. A developing trend in several large platforms is for user-to-user interactions to happen predominantly by sharing visual media content, like images and videos (e.g., Snapchat, Instagram, TikTok). Moreover, even platforms that are not primarily visual such as Facebook, WhatsApp, and Reddit, all offer media sharing capabilities. The recent years have demonstrated the large and ever-growing popularity of visual media. Users spend daily an average of *30 minutes* on Snapchat [31], exchange *95 million* photos and videos on Instagram, [32] and watch more than *1 billion* hours of video on YouTube [5].



(a) Open-Network **(b) Closed-Network**

Figure 1: (a) shows an example of open-network content. Here, content aims to please a large audience. In contrast, (b) shows an example of closed-network content, where the audience is one or few individuals who are familiar to the creator.

On most social platforms, users can share content with different target audiences in mind; settings commonly include *open-network* sharing (exposure to all other users on or off-platform), or *closed-network* sharing (exposure to a limited set of one or more of a users' friends) – Figure 1 shows an example. Despite these two distinct content creation and sharing modalities, past work in the data mining and adjacent communities has largely focused on content consumption predominantly in the open-network setting. Notably, the recommendations and information retrieval communities mainly tackle the problem of matching user preferences with openly accessible content (visual, or otherwise) like songs, videos, products, webpages, and other users via search and ranking paradigms. Additionally, many works in the computational social science domains focus on large-scale social dynamics and interaction processes such as virality, diffusion, and cascades of public information, tweets, media, and memes [2, 3, 8]. Comparatively, there is a gap in work on *content consumption* in closed-network settings where users often share

*Work done while author was on internship at Snap Inc.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411935

content to communicate personal moments, daily updates, inside jokes, and more [10]. Interaction with one's closest friends is vital to the end-user experience, making it a valuable research area.

Our work lies at the intersection of visual media and content consumption: specifically, we study consumption of visual media content in closed-network settings, where the content is created and shared between friends. Interestingly, the content shared in such networks could be especially meaningful or contextually relevant to the specified viewer audience given that the sender and viewer likely share an off-platform relationship of significance, in contrast to more general open-network consumption settings.

Given this nuance, our work aims to address key questions around *consumption of content* in *closed-network settings*. Specifically, we ask:

RQ1 Are there patterns in users' consumption of content with regard to the senders?

RQ2 Are these patterns associated with other social factors?

RQ3 Can we use social factors to predict future consumption intensity of the content sent by friends?

Our work makes contributions in response to these questions:

C1. We conduct the first empirical analysis of closed-network content consumption habits, using interaction data from Snapchat, a large-scale, multimedia-driven sharing service with over 200M active users. We show patterns in user's consumption (dwell time) behavior towards friends, and propose interpretable, parametric models based on log-normal decay to accurately capture this. Our models describe total and average dwell time behaviors across friends which outperform alternatives for 61.2% and 85.6% of users respectively (Section 4).

C2. In order to explain the consumption patterns, we investigate various social factors (notions of tie strength) that could be associated with preferentiality towards certain friends, and show that engagement-based factors strongly correlate with future consumption while network-based factors do so much less significantly. We also identify interesting social explanations which may underpin our correlational observations (Section 5).

C3. The correlation between consumption and social factors motivates us to study the task of ranking content with respect to their senders, which we call friend ranking. We show out-of-the-box learning-to-rank approaches can achieve strong predictive performance (0.815 P@1, 0.650 nDCG@10) in ranking friends based only on easily measured social factors, without the privacy or computational implications of any content-based factors, which is crucial in closed-network settings (Section 6).

2 RELATED WORK

We discuss related work in three adjacent spaces: user engagement modeling, social ties and influence, and user preference learning.

User engagement modeling. Many prior works tackled characterization, modeling, and forecasting specifically for explicit actions on public, open-network content such as page likes [26], tweets [11], follows [15], clicks [37] and general metrics [30]. However, owing to sparsity in explicit feedback, several recent works (including ours) consider implicit signals such as *dwell time* (time spent) on items in different engagement contexts. For example, [35] shows lognormal decay on user dwell times on short-text documents and uses them in collaborative filtering. Several works focus on dwell time in search contexts; [34] uses dwell time for candidate document reranking, while [6, 21] measure and model webpage dwell times using simple parametric forms. Most works do not consider dwell time in visual media; [16] is the closest, but focuses on open-network content. Several works consider community detection from engagement [1, 12, 27], but are not suited to content consumption. *Unlike all these works, ours (a) considers dwell time engagement with closed-network visual content, (b) identifies patterns in dwell time with respect to the senders, and (c) demonstrates its association with various social factors.*

Social ties and influence. Past work on measuring and quantifying social factors has been largely driven by the computational social science community, and mainly involves associating engagement with underlying network features. One line of work centers on tie strength and inference (link prediction). [25] utilizes network motifs on Twitter to identify strong ties. [33] proposes an unsupervised model to estimate tie strength from interactions. [19] learns a model to predict tie polarity using path, triad and signed edge features. Many works propose unsupervised network-based heuristics in link prediction settings to measure user-user affinity; [20] gives an overview. *Our setting differs in that we model engagement and consumption over a fixed set of links (friends), rather than inferring new ones.* Another line of work considers social influence on user behaviors. [28] shows positive correlation between users' chat frequency and similarity in their web searches. [3] demonstrates that a user's exposure to links shared by Facebook friends increases their own sharing propensity multifold. Social influence on user behaviors has also been explored in the context of open-network reshare cascades; [2] studies their predictability on Twitter given network features. Similarly, [8] shows that structural and poster/resharer features are predictive of cascade growth on Facebook. Recently developed social recommendation approaches [23, 29] leverage social relations in addition to public user-item feedback to improve recommendations. *In contrast, our work differs by (a) predicting future consumption from past engagement, and (b) focusing on closed-network phenomena.*

User preference learning. There is significant past work in learning user preferences of open-network content such as movies [14], documents [17], videos [4] and more. Much of this is driven by modern improvements in learning to rank (L2R) approaches from the information retrieval (IR) community [22]. While our work does not emphasize methodological novelty in this space, it has strong practical applications as we show later. *Our work (a) shows how a L2R paradigm can be adapted to rank friends of each user based on engagement, (b) shows relationships between consumption and social engagement, and (c) demonstrates strong performance without explicit content-based modeling, which is crucial in privacy-sensitive settings.*

3 BACKGROUND

3.1 Preliminaries

User designations. In each content sharing interaction, we refer to the source user as the *sender* and the recipient as the *viewer*. Two users may be senders and viewers for different interactions, but each interaction has a single notion of viewer and sender. When we refer to consumption behavior of a user, we consider interactions in which the *user* (viewer) views content shared by his/her *friends* (senders).

Measuring consumption. We use *dwell time* to measure the intensity of a consumption interaction (a single view). Dwell time is a powerful signal which can act as a proxy for a user's interest/preference; longer dwell times suggest continued interest. Although various contents can have varying durations, thereby influencing dwell times, [16] shows that the vast majority of views are quite short and unhindered by content duration – thus, we do not explicitly model this covariate.

Consumption settings. In this work, we refer to a user's consumption of content created by their friends and shared with them as *closed-network* consumption. Examples of closed-network consumption would be a user viewing a photo of a sunset shared by their partner, or of a user watching a video of a friend's child learning to swim. In contrast, we use *open-network* consumption to refer to the case in which consumed contents may be publicly accessible and created or shared without a constrained audience in mind. Examples of open-network consumption would be watching a YouTube video about cats, or watching a popular TV show. These designations imply different target audiences: in closed-network settings, content is *privately* created/shared with one or more social relations/contexts in mind.

3.2 Data Description

We study consumption behavior using data from Snapchat, a leading visual media content sharing platform with over 200M users. Our data includes consumption metrics (in terms of dwell times), as well as various social interactions. We introduce interaction types and collection process below.

Direct Snap. Users on Snapchat can share private image/video content samples called *Direct Snaps* with their friends in 1:1 or 1:few fashion. Direct Snaps (Snaps) are visible in the friend's inbox until watched. We capture the bidirectional edges between users and friends reflecting the number of Direct Snaps sent and received by each pair of users, and the associated dwell times.

Story. Users on Snapchat can post image/video content samples on their "My Story," which is a passive broadcast to their friends in a 1:all fashion – these are called *Story Snaps*. Story Snaps (Stories) are only visible for 24h, and not pushed to friends, but rather pulled (voluntarily). We capture the bidirectional edges between users and their friends reflecting the number of Story Snaps viewed for each of the two users forming an edge, as well as the associated dwell times.

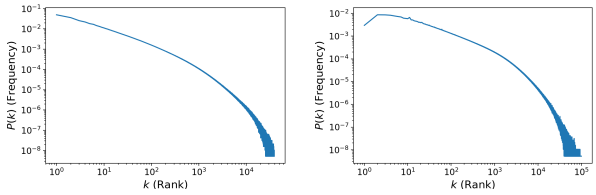
Chat. Users on Snapchat can also exchange private text messages called *Chats*, in a 1:1 fashion. We capture the bidirectional edges between users and their friends reflecting the number of Chats sent and received by each of the two users forming an edge.

Friend Graph. We additionally use an (undirected) graph which tracks friendship links between users, and time of link formation.

We collect data over the span of a consecutive two-month period in 2019; we impose different sampling constraints across months to support conditions to address different RQs while not processing extra data, and enable analysis of past engagement (1^{st} month) w.r.t future consumption (2^{nd} month). For the 2^{nd} month, we sample 13M unique viewers (users), with 29M unique senders (friends) and 976M directed interaction edges (488M undirected friendship edges), with many interactions per edge. Each viewer has $\geq$10 associated senders from whom he/she has viewed $\geq$50 Snaps. For the 1^{st} month, we additionally consider interactions between these 13M viewers and any senders from whom they viewed $\geq$1 Snap. Section 8.1 further details constraint rationale, data use in different experiments and collection considerations.

4 MODELING CLOSED-NETWORK CONTENT CONSUMPTION

Given these preliminaries, we begin by characterizing users' (viewers') consumption behaviors both in aggregate, and broken-down across friends (senders). Our work considers the consumption with respect to Snaps, the most private and common visual interaction.



(a) Media samples consumed **(b) Aggregate consumption time**

Figure 2: Rank-frequency plots of (a) media samples (Snaps) consumed, and (b) viewer time spent (note log scales). Consumption behaviors are highly skewed in closed-network content, demonstrating log-normal tendencies.

4.1 Aggregate consumption

We first consider overarching patterns across viewers' (users') aggregate consumption of content in terms of (a) number of consumed media samples (Snaps), and (b) amount of time spent on media samples in seconds.

Figure 2 shows the two phenomena in rank-frequency form, using consumption data from one (1^{st}) month of data. 2a shows the ratio of viewers who consumed k media samples. Likewise, 2b shows the ratio of viewers who spent k seconds consuming media samples. The artifact at the top-left of the plot is explained by discretization effects ($\sim$1 sec. is more common than $\sim$0). We note that both phenomena show a slow decline even in log-log scales, indicating their strong right-skewness and heavy tails. Empirically, we find that both are log-normally distributed. Formally:

*Definition 4.1 (Log-Normal Distribution (*LN*)).* Let T be a non-negative continuous random variable, such that $T \sim LN(\mu,\sigma)$. The PDF and CDF of T are given by:

$$f_{LN}(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\log t-\mu)^2}{2\sigma^2}} \qquad F_{LN}(t;\mu,\sigma) = \Phi\left(\frac{\log t-\mu}{\sigma}\right)$$

where $t \in (0,\infty)$, $\mu \in (-\infty,\infty)$ and $\sigma > 0$ are the mean and standard deviation of logT, and Φ indicates the standard normal CDF.

While log-normal decay is consistent with prior literature on various aggregate open-network consumption or engagement phenomena such as dwell time on public videos [16] and user voting behaviors on public news articles [18], our findings indicate that this behavior also holds in closed-network contexts.

4.2 Friend-specific consumption

Closed-network consumption occurs when users consume content shared by their friends, but how does this consumption happen? How do users apportion their consumption behaviors across friends, and are they partial to some friends more than others? Below, we focus on measuring consumption via two notions of intensity, reflecting "total" and "expected" behaviors with respect to each friend.

4.2.1 ***Total Consumption (TC).*** We describe total consumption intensity (TC) between a user and his/her friend as the total time spent by the user on consuming the friend's contents. Formally,

$$TC(u,f) = \sum_{c \in C_{f\to u}} \delta(c)$$

where $C_{f\to u}$ denotes the set of contents shared from friend f to user u, c denotes (without loss of generality) one such content sample, and $\delta(c)$ indicates the time spent by u in consuming the content c.

Note that TC is a *summation* of all the time spent on a given friend's contents, and is thus heavily mediated by the frequency of communication between two friends i.e., given equal propensity for user u to consume content samples from friends f_1 and f_2, then $TC(u,f_1) > TC(u,f_2)$ if $|C_{f_1 \to u}| > |C_{f_2 \to u}|$. Thus, given a supposed asymmetry in communication frequency across friends, TC should be affected. Figure 3 demonstrates that this is indeed the case, showing patterns in TC for several randomly sampled users with 100 friends – in all subplots, the y-axis indicates a user's normalized consumption times over friends, and the x-axis indicates the friends ranked from highest to lowest TC. We find that TC is highly skewed, such that a few friends dominate the metric, though the degree of dominance by top friends differs. The differences in y-axes values in the figures indicate that the skewness can vary considerably due to the fact that different users demonstrate different preferences and communication habits towards their friends.

We ask, is there a pattern in the decay of TC over a user's friends? To study this, we evaluate how well TC habits of many real users are described by parametric forms. We empirically analyzed viewing behaviors of 100K sampled users with ≥50 friends from our base dataset. We used a sample for computational efficiency, and a larger friend/sender threshold to better capture distribution tails.

Specifically, we aimed to describe the rank-intensity distributions (as in Figure 3) for each of the viewers as one of several parametric forms. Although the underlying distributions we consider are discrete, we use continuous approximations in line with past literature when handling large outcome spaces [13]. We considered the truncated (T) variants of seven underlying candidate distribution types: Exponential (E), Log-logistic (LL), Inverse Gaussian (IG), Weibull (W), Gamma (G), Pareto (P) and Log-normal (LN), which have demonstrated effectiveness in prior dwell time literature [9, 16, 21, 38]. We optimized these parametric forms for each user via maximum likelihood estimation (MLE) to achieve the best fit in each case, and measured the performance of these distributions by the fraction of users which were best-modeled (highest likelihood) by each of them versus the others (in one vs. rest fashion). Additionally, since the distribution for different users depends on number of friends (maximum rank), we evaluate goodness-of-fit with respect to right-truncated distributions, where the probability measure is defined as 0 for values larger than the true number of friends n (wlog), and the remaining density on (0,n] is renormalized to integrate to unity. Table 1 shows that the TLN (truncated log-normal) form strongly and consistently outperforms alternatives – the table is meant to be read as *"TLN outperforms TE for 89.7% of users."* We define the TLN formally as:

*Definition 4.2 (Truncated Log-Normal (*TLN*) Distribution).* Let T be a non-negative continuous random variable, such that $T \sim TLN(\mu,\sigma,n)$. The PDF and CDF of T are given by:

$$f_{TLN}(t;\mu,\sigma,n) = f_{LN}(t;\mu,\sigma) \Big/ Z \qquad F_{TLN}(t;\mu,\sigma,n) = F_{LN}(t;\mu,\sigma) \Big/ Z$$

where $t \in (0,n]$, $Z = F_{LN}(n;\mu,\sigma)$ (truncation normalization constant) and μ,σ match Defn. 4.1.

Dashed red lines superimposed on the subplots in Figure 3 indicate the MLE fits for the TLN distribution in each case, closely following the user behaviors despite individual variations.

4.2.2 ***Expected Consumption (EC).*** We describe expected consumption intensity (EC) between a user and his/her friend as the expected (average) time spent by the user on consuming the friend's shared contents. Formally,

$$EC(u,f) = \sum_{c \in C_{f \to u}} \delta(c) \Big/ |C_{f \to u}|$$

where the symbols are consistent with the TC setting, and the $|\cdot|$ operator denotes set cardinality.

Unlike TC, EC is an average or *expectation* of time spent on a given friend's contents, and the normalization of the communication frequency between two friends. The EC represents the expected amount of time spent on a single content sample from the friend. The EC is arguably a more interesting measure to study since it can be used to reason about the attention a user is willing to spend on his/her friend *given the opportunity*. This is a key consideration, as some of a user's friends may communicate more or less frequently, giving a user more or less opportunity to consume content. Skewness in communication frequencies is a reality in social communications, and thus produces skewed opportunities for consumption. But when normalizing for this skewness via EC, we are interested to discover how consumption is distributed? Are friends equally likely to spend time on each other in this setting, or is there latent preferentiality when choosing to consume friends' content given the opportunity?

Figure 4 shows EC patterns for several randomly sampled users with 100 friends (EC estimated across ≥50 content samples per friend) – the subplots indicate normalized EC on the y-axis and friend rank on

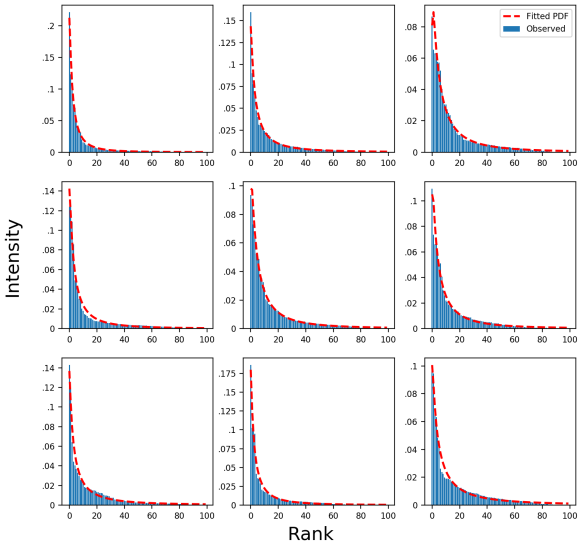


Figure 3: Rank-intensity plots of *total* consumption (TC) partitioned over friends for 9 sampled viewers (note linear scale). TC is highly skewed over top friends, and closely follows our proposed TLN distribution (dashed red lines).

Table 1: Performance comparison for total consumption (TC) and expected consumption (EC) under MLE log-likelihood: log-normal forms win (⋆), outperforming alternatives by large margins.

	TE	TLL	TIG	TW	TG	TP	TLN
TC	89.7%	92.6%	61.2%	84.7%	85.6%	100.0%	⋆
	UTE	**UTLL**	**UTIG**	**UW**	**UTG**	**UTP**	**UTLN**
EC	98.0%	95.3%	85.6%	97.3%	97.2%	99.7%	⋆

the x-axis, similarly to Figure 3. Note that no preferentiality would be indicated by a uniform distribution. Fascinatingly, the results show strong non-uniformity over the friends, given that users' EC tends to be strongly skewed at the head of the distribution and implies significant preferentiality towards certain friends. Our analysis across many users shows a variably placed, but consistently steep dropoff of EC over the top 1 or few friends, with an eventual taper tending towards uniformity in the tail (especially apparent in Figure 4 where the tail is emphasized). This suggests that even when accounting for different communication frequencies and consumption opportunities from their friends, users are more likely to prioritize and pay attention to some of their friends' contents more than others'. However, there is likely a minimal, or token, EC afforded to content from more casual or distant friends (towards the right on the x-axis).

Interestingly, the tails of the EC distributions in Figure 4 are rather fat, and seem to not decay significantly or at all after a certain friend rank (depending on the user), which is consistent with our hypothesized token EC afforded by the user to their friends (independent of rank). To model this, we adopt the same MLE-based experimental setting to evaluate goodness-of-fit, but we consider the aforementioned truncated distributions *mixed* with an additional uniform component to capture this supposed token EC. The uniform component describes a token EC allotted to all friends, which could intuitively correspond to a level of baseline interest or time spent by the user in consuming content, anything over which is described according to a more flexible (non-uniform) decay process.

We empirically analyzed viewing behaviors of 100K randomly sampled users with ≥50 friends, where each user had viewed ≥50 content samples from each friend. We used this minimum content samples designation to enforce robustness in the EC estimate, since an average over few samples may be unreliable. Table 1 shows results (uniform mixtures indicated by a U prefix), indicating that the UTLN (uniform truncated log-normal mixture) strongly outperforms alternatives. We define the UTLN formally as:

*Definition 4.3 (Uniform Truncated Log-Normal Mixture (*UTLN*) Distribution).* Let T be a non-negative continuous random variable, such that $T \sim \text{UTLN}(\mu,\sigma,n,\vartheta)$. The PDF and CDF of T are given by:

$$f_{\text{UTLN}}(t;\mu,\sigma,n,\vartheta) = \vartheta f_U(t;n) + (1-\vartheta) f_{\text{TLN}}(t;\mu,\sigma,n)$$
$$F_{\text{UTLN}}(t;\mu,\sigma,n,\vartheta) = \vartheta F_U(t;n) + (1-\vartheta) F_{\text{TLN}}(t;\mu,\sigma,n)$$

where $\vartheta \in [0,1]$ (mixture probability), $f_U(t;n) = 1/n$ (uniform PDF), $F_U(t;n) = t/n$ (uniform CDF) and t,n,μ,σ match Definition 4.2.

Again, dashed red lines superimposed on the subplots in Figure 4 indicate the impressive closeness of the MLE fits for the UTLN distribution in each case. Note the flexibility of the mixture to well-model the nuances of the distributions.

Overall, we show that most users tend to be at least somewhat preferential towards certain friends, with few users being extremely so. We now turn to quantify this preference by computing the Kullback-Leibler (KL) divergence between the expected consumption distribution and uniform distribution (assuming consumption time is distributed equally among friends) for each user. We call the obtained KL divergence as "preferentiality score". Figure 5a plots the cumulative distribution of the preferentiality scores across all users. Non-zero score indicates a positive degree of preferentiality, and we observe that most users meet this criteria. Most users show a preferentiality akin to that in bucket 1 (top friend with 2× EC of the next), but there are also some users who are extremely preferential

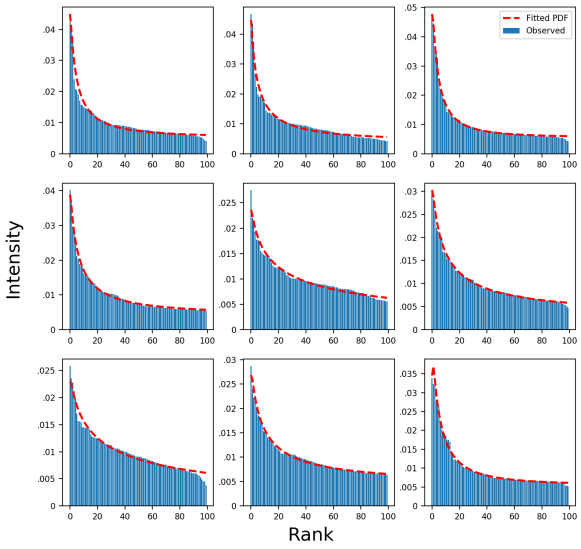


Figure 4: Rank-intensity plots of *expected* consumption (EC) partitioned over friends for 9 sampled viewers (note linear scale). EC is highly skewed over top friends, after which decay slows considerably – note the fat tails. The pattern closely follows our proposed UTLN mixture distribution.

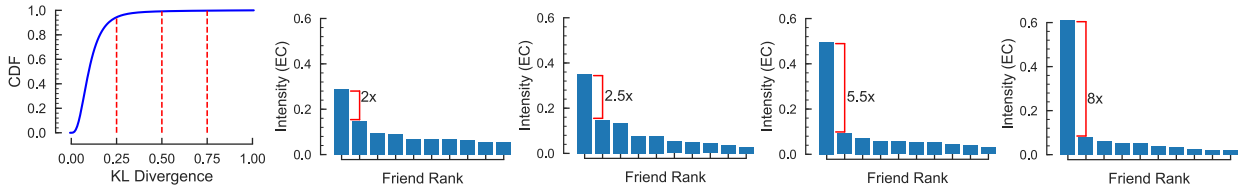
as in bucket 4 (top friend with 8× EC of the next). Next, we consider if this preferentiality is likely to be socially motivated.

5 ASSOCIATION WITH SOCIAL FACTORS

We found that users are preferential towards certain friends when consuming their content. Therefore, we hypothesize that content consumption is mainly driven by two phenomena: (a) social factors, and (b) content-based factors. Social factors might indicate the preference of a user to engage with a friend's content due to the underlying relationship (a user may not care for the content itself, but care for the sender). Conversely, content-based factors might indicate user interests in various types of contents (some users might prefer videos about cars, others might prefer videos about dogs). In practice, these factors are impossible to disentangle; it is not possible to counterfactually evaluate how a user would consume the same content from two different friends at the same time. Moreover, content-based factors are extremely challenging to quantify in closed-network settings, given both subjectivity in user interests as well as privacy/encryption of content (making inspection and analysis of content infeasible). On the other hand, social factors in association with consumption are appealing to study, given (a) prior literature on measuring and quantifying social factors (i.e. tie strength) between individuals in network contexts [24], (b) no need to view private content, and (c) lighter computational load compared to deep/visual content-based methods. But are social factors actually significantly associated with consumption behaviors? We investigate the relationship in this section.

We aim to understand the correlations between *past social factors* and *future consumption behaviors* between users and friends. We consider various tie strength notions between a user/viewer u and friend/sender f given their engagement and network-based relations in the past:

- Chat (direct textual content to friends) engagement-based
  - Chats sent by u to f
  - Chats received by u from f



(a) KL CDF v.s uniformity **(b) Bucket 1** **(c) Bucket 2** **(d) Bucket 3** **(e) Bucket 4**

Figure 5: Most users are preferential w.r.t expected consumption (EC), with few, select friends receiving markedly higher attention than others. (a) shows the empirical CDF of user preferentiality, taken over all 13M users towards their friends, where preferentiality is measured by KL divergence of a user's EC distribution w.r.t. a uniform one. (b)-(e) show EC distributions for 4 sample users from each of the buckets designated in (a) – note the increasing degrees of preferentiality across buckets. Most users fall in the 1^{st} bucket, where preference is marked but not extreme, e.g. the top friend has $2\times$ the EC of the next in (b). However, some users are extremely preferential, e.g. the top friend has $8\times$ the EC of the next in (e).

- Snap (direct visual content to friends) engagement-based
  - Snaps sent by u to f
  - Snaps received by u from f
  - Incoming EC time to u's Snaps by f
  - Outgoing EC time to f's Snaps by u
- Story (broadcast visual content to friends) engagement-based
  - Stories posted by u and viewed by f
  - Stories viewed by u and posted by f
  - Incoming EC time to u's Stories by f
  - Outgoing EC time to f's Stories by u
- Network-based (features about network context)
  - Common neighbors between u and f
  - Friendship tenure between u and f (time since tie formation)

To measure the relationship strength we used *Spearman rank correlation* [36], which evaluates the strength and direction of the relationship between two variables. It is based on the ranks, rather than raw values, between the two variables. Given user u with n friends (wlog), we consider a past social factor $\vec{s}$ and future consumption rank vector $\vec{c}$, such that $\vec{s}_i$,$\vec{c}_i$ are the ordinal ranks of the i^{th} friend (largest to smallest). The Spearman correlation for u is defined as $\rho_u = 1 - \left(6\sum_i d_i^2\right) / \left(n\left(n^2-1\right)\right)$ where $d_i = c_i - s_i$. ρ_u is defined on $[-1,1]$, with 0/-1/1 implying no/perfectly -ive/perfectly +ive correlation, respectively. Significance can be computed using permutation tests which give p-values against the null hypothesis H_0, which supposes that the two variables are uncorrelated.

The left side of each subfigure in Figure 6 shows empirical CDFs of ρ_u across all viewers under various notions of past tie strength s. Blue curves which are right-skewed imply that most users demonstrate a +ive ρ_u between the given $\vec{s}$ and $\vec{c}$ (EC), left-skewed imply the -ive case, and the dashed red curve indicates the "null" scenario (i.e. all users exhibiting no association). Since significance is impacted by sample size (in this case, friend count), we show ratios of users whose ρ_u had associated significance $p < .05$ with different friend counts in the right side of each subfigure (blue/red curves show +/-ive correlation, respectively). We see that these ratios become more stable when considering more friends, and +ive correlations become more pronounced. Altogether, Figure 6 shows several interesting findings, which we discuss below.

Strong +ive correlation with engagement. Firstly we can see the majority of users fall in the first group (+ive correlation) for all engagement-based tie-strength measures (a)-(j). For example, future EC is +ively correlated for 88%/95%/83% of users for number of Chats sent (a), outgoing Snap EC (g), and outgoing Story EC (i). In all cases, considering more friends over measurement of ρ_u increases significance. For example, the right-hand side of (a) shows that over 90% of users with 100 friends have significant +ive correlations between number of Chats sent and future consumption.

Directional importance. Secondly, we can notice that while *both outgoing and incoming* metrics are +ively correlated with future EC intensity, outgoing metrics show slightly to markedly stronger correlations to their incoming counterparts for all 3 engagement metrics. For example, over 83% of users have +ive ρ_u in outgoing Story EC (i), compared to 72% in incoming Story EC (j). This is intuitive, since the perception of any relationship (and propensity to engage) between u and f likely differs between the two. But, as one might expect, u's outgoing (voluntary) engagement with f tends to better capture his/her affinity for f compared to the inverse (involuntary) setting.

Consistence with past preferences. Thirdly, of all factors, past outgoing Snap EC is most strongly correlated to future Snap consumption EC (g). The past being predictive of the present is especially meaningful in this setting, as it shows that users' time-spending preferences across friends are not too different *despite entirely different content being viewed during past and future periods*; the correlation strength suggests that a large portion of the variance between past and future consumption is explained by the underlying social context of the interactions. This suggests that content-based factors may offer limited added value.

Target audience and interaction depth. Fourth, we note that Story engagement (e-f, i-j) does have a consistent +ive association with future consumption, but is markedly less than Snap/Chat measures; this is likely due to Story engagement carrying less personalized social context (since it is broadcast to all friends), in contrary to the more private nature of Snap/Chat engagement. Moreover, between Snap/Chat settings, we notice that despite Snap EC features (g-h) demonstrating very strong correlations, past Chat frequency (a-b) correlates more strongly with future consumption than Snap frequency (c-d), likely due to the different social contexts of private 1:1 text messaging, versus possible 1:few Snap messaging. Intuitively, a user might casually share dinner photos and flattering selfies with a few friends, but only initiate conversation (Chat) with those who they expect responses from.

Lower correlation with network-based factors. We notice considerably lower and less strong association between network-based features like common friends (k) and friendship tenure (l) with future EC. The former (k) shows an interesting complement to traditional link prediction literature, which holds that common neighbors are very useful as a tie-strength measure, albeit in the context of

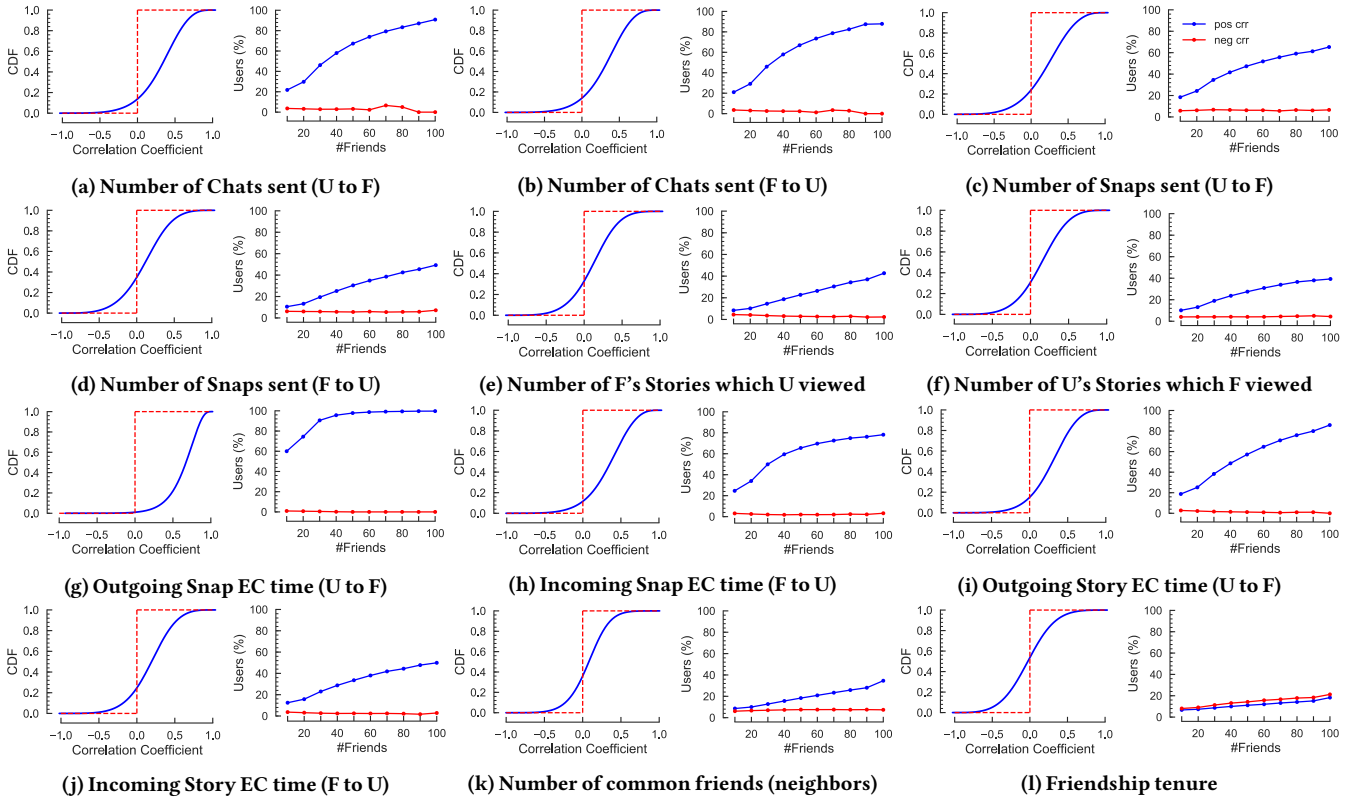


Figure 6: Consumption is significantly correlated with engagement-based social factors. Each subfigure shows the (blue) CDF of Spearman ρ between a *past-measured social factor* and *future-measured EC* taken over friends of each user (left), and the ratio of +/-ive significance ($p < .05$) when considering users with varying friend counts (right). Right-skewed CDFs (left) show that more users have +ive correlation between past engagement and future EC (88% in (a), 78% in (c), 90% in (h) etc.), with ratio of +ive significance (blue, right) clearly substantiating the effect as we consider users with more friends (over 90% users w/ ≈100 friends have significant +ive correlations). Network-based features like common neighbors and tie duration are markedly less clearly correlated, with 62% +ive for (k) and 48% +ive for (l).

link formation. (k) suggests that once these ties are formed, common neighbor features are still +ively associated (for 62% of users) with EC, but much less so compared to other engagement-related features. The latter suggests that contrary to expectations, users overall do not convincingly prefer very old or new friends in terms of future consumption (48% +ive, 52% -ive). We hypothesize that highly preferred friends may be a mix of the two; long-standing friends may share content better imbued with relationship and meaning, but may not be as immediately relevant to a user's current social situation (consider a user with old friends from their childhood, but with new friends in a new workplace).

Altogether, our results suggest that many social factors are considerably correlated with future consumption, even in the absence of explicit content modeling. This offers significant promise in solving predictive modeling problems around closed-network user/content prioritization, which are common in many social platforms.

6 PREDICTING FUTURE CLOSED-NETWORK CONTENT CONSUMPTION

The correlations we reported lead us to ask, how well can we *predict* future consumption using social factors? In closed-network content sharing modes, users may receive multiple (possibly visual) messages and must choose which users' contents they want to consume (typically given intrinsic time and interest budgets) through an interface which gates the content behind an inbox from each friend (e.g. Figure 7). Moreover, content from any of these users may be (un)available at different times, leading to ranking over subsets of these friends, and indirectly their contents. Thus, it is important to make fast, on-the-fly choices to feature the most relevant and engaging content from a user's friends. Solving this task can lead to improved prioritization of content, increased engagement and time spent, and improved user experiences.

Friend ranking. Inspired by the notion of Learning to Rank (L2R) in the IR community, typically considered on documents in search contexts, we propose a *friend ranking* setup where instead of a query q and candidate documents $\mathcal{D}$, we have a user u and friends $\mathcal{F}$. The goal is to learn a function which produces a ranking which favors f_i to f_j if u prefers f_i to f_j (wlog) in terms of consumption, on unseen data. As L2R is a well-studied problem in other contexts, we adapt and consider existing approaches for our setting. We pre-process training data for friend ranking using large-scale user/friend interaction data, with past engagement/ tie strength scores between u and f

(wlog) as features over one month of data, and defining labels using future consumption metrics from the next month. We emphasize that our goal here is not to contribute a new model to L2R literature, but rather (a) evaluate predictive capacity of our results on social significance and friend-based preferentiality in closed-network consumption, and (b) demonstrate a novel and practical application of L2R and associated findings.

Defining Labels. L2R algorithms in retrieval tasks are often based on a dataset of queries documents with well-defined relevance labels (i.e. 1 to 5). However, in our setting, ground-truth labels indicating well-defined consumption preference are not explicitly available. Thus, we use the EC time as the proxy to gauge users' preference towards their friends (motivated by the "expected time spent when given the opportunity" setting), using empirical ECs estimated over all Snaps viewed by u from f. One important consideration is that users may have different consumption patterns, meaning that two individuals may dwell differently on content sent by their most-preferred friends (i.e. Figure 4). Thus, direct use of raw EC as preference labels in training would bias against viewers who spent shorter time on their friends' content. To alleviate this, we min-max normalize friends' ECs for each user into [0,1] to get preference scores to use as ground-truth labels.

Evaluation Setup. We report three retrieval metrics to evaluate the effectiveness of the various friend ranking setups. The first metric is precision over top k friends (reported as P@[1,3,5]), classically used for retrieval tasks with binary labels; we adapt for the continuous label case by defining the top-5 preferred friends as relevant and the rest as irrelevant. Next, we report normalized discounted cumulative gain at different ranks (reported as nDCG@[1,3,5,10]), which emphasizes ranking order rather than only discovery. It is defined as nDCG@k = DCG@k/IDCG@k where $\text{DCG@k} = \sum_{i=1}^{k}(2^{l_i}-1)/(\log_2(l_i+1))$ and $\text{IDCG@k} = \sum_{i=1}^{k}(2^{i}-1)/(\log_2(i+1))$, and l_i denotes the ground-truth position of the friend predicted at position i. Finally, we consider a winner-takes-all (WTA) measure which is defined as WTA = 1 if the top preferred friend is retrieved correctly, else WTA = 0. WTA is motivated by correctly inferring the strong preferentiality usually observed towards the best friend (i.e. Figures 4-5). All reported metrics are averaged across users. We split the dataset into train, test, and validation sets. Each dataset spans 100K unique viewers and ≈2.5M pairs of viewer-sender relationships.

Learner choices. We evaluate friend ranking performance using different methods. We first consider an unsupervised method which simply sums all feature values and ranks friends based on that score. We also consider several pointwise learned approaches including linear regression, gradient boosted trees, and random forest, trained to predict consumption intensity. We also consider a high-performant pairwise method, LambdaRank [7]. Results are in Table 2. We observe that unsupervised and linear methods perform considerably worse across all metrics compared to nonlinear methods. Since LambdaRank performs best, we use it in further experiments.

Feature importance. We consider the impact of past social factors in ranking performance via feature ablation – see Table 3. We observe that (a) private visual communication (Snap) features have surprisingly strong importance even in isolation, and (b) of the other two engagement-based factors, Chat is more predictive than Story, substantiating our observations on consistence with past preferentiality and target audience from Section 5. Chat & Snap together outperform Snap & Story, further demonstrating effects of interaction

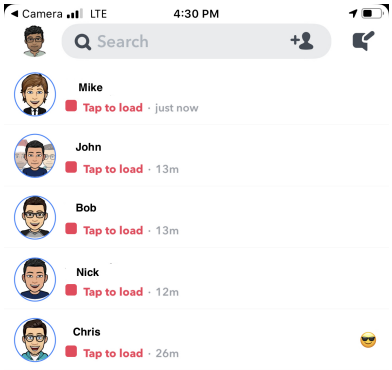


Figure 7: An example of closed-network content prioritization on Snapchat: a user must choose between viewing unseen contents sent by 5 friends, and if, when, and how long they want to consume it; ranking can improve user experience and engagement propensity.

Table 2: Nonlinear ranking methods strongly outperform simpler (linear and naïve unsupervised) methods, in both pointwise and pairwise EC-based friend ranking.

	Precision@k			nDCG@k				WTA
	1	3	5	1	3	5	10	1
Unsupervised	0.617	0.560	0.504	0.271	0.334	0.387	0.476	0.162
Linear Regression	0.613	0.552	0.505	0.265	0.331	0.389	0.487	0.158
Gradient Boosting	0.811	0.724	0.643	0.451	0.515	0.568	0.646	0.311
Random Forest	0.813	0.724	0.644	0.455	0.518	**0.571**	**0.650**	**0.316**
LambdaRank	**0.815**	**0.725**	**0.645**	**0.456**	**0.519**	**0.571**	**0.650**	**0.316**

Table 3: Engagement-based social factors are much more predictive in learned friend ranking than network-based ones.

	Precision@k			nDCG@k				WTA
	1	3	5	1	3	5	10	1
Chat	0.613	0.538	0.489	0.250	0.309	0.364	0.463	0.143
Snap	0.803	0.719	0.643	0.452	0.517	0.571	0.651	0.315
Story	0.537	0.485	0.450	0.209	0.272	0.327	0.430	0.116
Chat & Snap	0.813	**0.725**	**0.646**	0.455	**0.519**	**0.572**	**0.650**	**0.316**
Chat & Story	0.625	0.501	0.411	0.260	0.322	0.378	0.475	0.150
Snap & Story	0.804	0.717	0.641	0.451	0.516	0.570	0.648	0.313
All Engagement-based	**0.814**	**0.725**	**0.646**	0.455	**0.519**	**0.572**	**0.650**	0.315
Common Neighbors	0.400	0.368	0.352	0.133	0.181	0.228	0.327	0.065
Friendship Tenure	0.514	0.430	0.389	0.215	0.252	0.295	0.390	0.128
All Network-based	0.424	0.382	0.361	0.155	0.203	0.249	0.348	0.084
All	**0.815**	**0.725**	0.645	**0.456**	**0.519**	0.571	**0.650**	**0.316**

Table 4: Learned EC-based friend ranking strongly outperforms unsupervised network-based link prediction heuristics in ranking friends according to future consumption.

	Precision@k		nDCG@k		WTA
	1	5	1	5	1
Common Neighbors Index	0.233	0.193	0.078	0.127	0.038
Inverse Friend Degree Index	0.175	0.180	0.059	0.116	0.030
Adamic-Adar Index	0.252	0.212	0.080	0.134	0.037
Resource Allocation Index	0.259	0.218	0.082	0.138	0.038
Preferential Attachment Index	0.180	0.178	0.059	0.113	0.029
Jaccard Index	0.266	0.219	0.086	0.141	0.040
Salton Index	0.270	0.222	0.087	0.143	0.041
Sørensen Index	0.266	0.219	0.086	0.141	0.040
Hub Promoted Index	0.253	0.220	0.084	0.142	0.040
Hub Depressed Index	0.256	0.215	0.083	0.138	0.039
Local LHN Index	0.244	0.215	0.082	0.140	0.040
Learned friend ranking	**0.815**	**0.645**	**0.456**	**0.571**	**0.316**

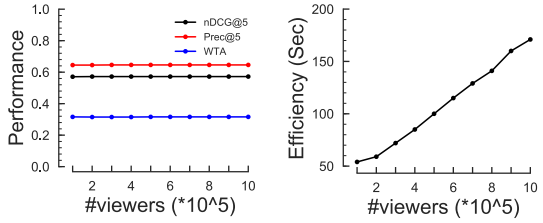


Figure 8: Friend ranking shows (a) stable performance, and (b) linear training scalability in number of viewers.

depth. Grouping all engagement-based factors outperforms exclusion of any one of them. Network-based features are comparatively less effective; friendship tenure and common neighbors are less predictive than any engagement-based feature, and the combination actually hurts performance of friendship tenure alone. Comparable results between only engagement-based and all factors (rows 7 and 11) shows that network-based features do not convincingly improve performance when considered in addition to engagement-based ones.

Comparison with link prediction methods. Most prior work on ranking users considers user-friend link prediction applications using network-based features [24]. The key idea is that users prefer other users who are close friends-of-friends. Though our friend ranking task is markedly different than link prediction due to the unique context of (a) consumption-based preference, and (b) re-ranking existing links rather than ranking possible new ones, we evaluate relative performance of strong link prediction heuristics in our task. Table 4 shows performance of these methods (see Section 8 for their definitions) compared to the learned friend ranking solution. We see that traditionally understood network-based link prediction features are much less effective in predicting future EC compared to a learned ranking. This echoes our intuition from Section 5 which suggests that consumption-based friend preference is largely associated with *engagement-based* rather than *network-based* factors. This finding illustrates that ranking for *friendship creation* and *friendship time-spending/attentive preference* are markedly different.

Sensitivity to data size. We evaluate sensitivity with respect to data size in the friend ranking task in terms of both accuracy and scalability. Figure 8(a) shows that using various data sizes (from 100K viewers/2.5M friends to 1M viewers/25M friends) gives fairly consistent results for all metrics, suggesting a simple, but high-performant ranker can be built with comparatively little user-friend data. Figure 8(b) demonstrates linear scaling in friend ranking model training with respect to number of viewers, demonstrating fast and efficient model learning using social features.

7 CONCLUSION

We conducted the first large-scale study of content consumption behavior in a closed-network setting, using a dataset of over 13M viewers, 29M senders, 966M edges and billions of fine-grained interactions from Snapchat. Our analyses produce several interesting academic and application-oriented findings: We discover clear patterns and propose models for describing closed-network consumption habits in terms of total and expected consumption (TC, EC) across friends; our models (TLN, UTLN) consistently outperform next-best alternatives in modeling user behaviors (for 61.2%, 85.6% of users respectively). We show that users are preferential, spending much more time on content shared from certain friends than others. We next study associations between various social factors, or notions of tie-strength, based on engagement-based and network-based features from (past) user-friend interactions and (future) consumption behaviors. Our results demonstrate that consumption-based preferentiality (a) is strongly correlated with engagement-based factors, (b) differs in correlation strength based on engagement directionality, (c) is strongly consistent with past-exhibited preferences despite natural differences in underlying content shared in past and future contexts, (d) is influenced distinctly by different target audiences and interaction depths of engagement habits, and (e) is considerably less correlated with network-based social factors, which shows an interesting difference to link prediction scenarios. Finally, we show that social factors can be leveraged for ranking friends on social platforms in closed-network settings using out-of-the-box L2R methods, demonstrating strong performance (0.815 P@1, 0.650 nDCG@10) by themselves while incurring neither the privacy nor computational cost of explicit content-based preference modeling.

REFERENCES

[1] Majid Alfifi, Parisa Kaghazgaran, James Caverlee, and Fred Morstatter. 2019. A large-scale study of ISIS social media strategy: Community size, collective influence, and behavioral impact. In *ICWSM*.
[2] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *WSDM*.
[3] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *www*. ACM.
[4] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent cross: Making use of context in recurrent recommender systems. In *WSDM*.
[5] YouTube Offical Blog. 2017. You know whatś cool? A billion hours.
[6] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A context-aware time model for web search. In *SIGIR*. ACM.
[7] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* (2010).
[8] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *WWW*.
[9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* (2009).
[10] Hana Habib, Neil Shah, and Rajan Vaish. 2019. Impact of Contextual Factors on Snapchat Public Sharing. In *CHI*. ACM.
[11] Liangjie Hong, Ovidiu Dan, and Brian D Davison. 2011. Predicting popular messages in twitter. In *WWW*.
[12] Parisa Kaghazgaran, James Caverlee, and Anna Squicciarini. 2018. Combating crowdsourced review manipulators: A neighborhood-based approach. In *WSDM*.
[13] Gerald Keller. 2015. *Statistics for Management and Economics, Abbreviated*. Cengage Learning.
[14] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *KDD*.
[15] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *WWW*.
[16] Hemank Lamba and Neil Shah. 2019. Modeling Dwell Time Engagement on Visual Multimedia. In *KDD*. ACM.
[17] Dik L Lee, Huei Chuang, and Kent Seamons. 1997. Document ranking and the vector-space model. *IEEE software* (1997).
[18] Kristina Lerman and Rumi Ghosh. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *ICWSM*.
[19] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting positive and negative links in online social networks. In *www*. ACM.
[20] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *JASIST* (2007).
[21] Chao Liu, Ryen W White, and Susan Dumais. 2010. Understanding web browsing behaviors through Weibull analysis of dwell time. In *SIGIR*. ACM.
[22] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* (2009).
[23] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. 2011. Recommender systems with social regularization. In *WSDM*. ACM.
[24] Víctor Martínez, Fernando Berzal, and Juan-Carlos Cubero. 2017. A survey of link prediction in complex networks. *CSUR* (2017).
[25] Rahmtin Rotabi, Krishna Kamath, Jon Kleinberg, and Aneesh Sharma. 2017. Detecting strong ties using network motifs. In *www*. ACM.
[26] Nikos Salamanos, Elli Voudigari, Theodore Papageorgiou, and Michalis Vazirgiannis. 2012. Discovering correlation between communities and likes in facebook. In *ICGCC*. IEEE.
[27] Neil Shah, Danai Koutra, Tianmin Zou, Brian Gallagher, and Christos Faloutsos. 2015. Timecrunch: Interpretable dynamic graph summarization. In *KDD*.
[28] Parag Singla and Matthew Richardson. 2008. Yes, there is a correlation:-from social networks to personal behavior on the web. In *www*. ACM.

[29] Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu. 2013. Exploiting local and global social context for recommendation. In *IJCAI*. AAAI.
[30] Xianfeng Tang, Yozen Liu, Neil Shah, Xiaolin Shi, Prasenjit Mitra, and Suhang Wang. 2020. Knowing your FATE: Friendship, Action and Temporal Explanations for User Engagement Prediction on Social Apps. In *KDD*.
[31] TIME. 2016. Heres How Much Time Snapchat Users Spend on the App. http://time.com/4272935/snapchat-users-usage-time-app-advertising/.
[32] TIME. 2016. Instagram Just Hit the 500 Million User Mark. http://time.com/money/4376329/instagram-users/.
[33] Rongjing Xiang, Jennifer Neville, and Monica Rogati. 2010. Modeling relationship strength in online social networks. In *WWW*. ACM.
[34] Songhua Xu, Hao Jiang, and Francis Chi-Moon Lau. 2011. Mining user dwell time for personalized web search re-ranking. In *IJCAI*. AAAI.
[35] Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *KDD*. ACM.
[36] Jerrold H Zar. [n.d.]. Spearman rank correlation. Wiley Online Library.
[37] Yuchen Zhang, Weizhu Chen, Dong Wang, and Qiang Yang. 2011. User-click modeling for understanding and predicting search-behavior. In *KDD*. 1388–1396.
[38] Yuan Zhang, Tianshu Lyu, and Yan Zhang. 2018. Cosine: Community-preserving social network embedding from information diffusion cascades. In *AAAI*.

8 APPENDIX

8.1 Environment Details

Data Collection. We extract relevant user engagement and friend graph data discussed in Section 3.2 from Snapchat cloud-based relational data storage using SQL. We do not impose any constraints on users for candidacy in our base dataset beyond those which are required to study time-spending and preferentiality behaviors across friends. We collect data from two months. For the 2nd month, we consider users who have $\geq$10 friends who sent them $\geq$50 Snaps each during the month. From these, we sample 13M unique viewers, and consider the 29M unique senders and 976M (sender, viewer) and (viewer, sender) edges between them, each of which has numerous features related to engagement-based and network-based interactions. For the previous (1st month), we collect engagement data between each of these 13M unique viewers and each sender who sent $\geq$1 Snap. The different constraint for the 1st month is set to study TC across all senders (even for those who do not send frequently). We use these two months of data to model user consumption behaviors, correlate past social factors (1st month) with future consumption behaviors (2nd month), and predictive modeling. Since most of our analysis is viewer-centric, we often call viewers the "users" and senders their "friends." From this underlying data, we sample further for different analyses.
Software Environment. Most analysis was done using a Python 3 kernel in Jupyter Notebooks. For consumption behavior modeling in Section 4, we utilized the `statsmodels` library, and specifically the `GenericLikelihoodModel` module, which allows flexible specification of likelihood functions, constraints, and optimization routines. We optimized likelihood using Nelder-Mead optimization. The underlying implementations for several common PDF/CDFs were pre-defined in the `SciPy` library's `stats` module. During fitting, we fixed location to 0 since consumption (dwell times) are nonnegative.

For correlative analysis and significance testing in Section 5, we again used the `SciPy` library's `stats` module's built-in functionalities. For both Section 4 and 5, we used Python's `multiprocessing` library, taking advantage of data parallelism inherent in the tasks to distribute batches of user analysis tasks to different cores. This led to more time-efficient analysis.

For predictive analysis and ablation studies for friend ranking in Section 6, we used the `scikit-learn` library implementations of `LinearRegression`, `GradientBoostingRegressor` and `RandomForestRegressor`, and the `lightgbm` library implementation of LambdaRank-based optimization. We computed link prediction heuristics (Section 6) using a variant of SQL on structured friend graph data stored in a relational database.
Hardware. We use a single Google Cloud Platform *n1-standard-96* (96 vCPUs, 360 GB RAM) virtual machine. High vCPU count enables faster evaluation on larger datasets for some of our easily data-parallel operations like parametric modeling and correlation computation. Large RAM simplifies data loading operations on many user interactions in memory simultaneously. However, this analysis is doable on commodity hardware with more efficient implementations (using incremental data loading) and more time (using less parallelization).

8.2 Link prediction heuristics used

Link prediction heuristics have been used historically to evaluate the propensity of a nonexistent friendship/tie to form in an underlying friend graph; thus, they can be seen as a network-based notion of friend rank. [24] gives an overview of such metrics and their definitions are below. They depend on node neighborhoods and degrees between a user u and friend f. We use N(u) to denote the neighbors of user u in the (undirected) friend graph.

- **Common Neighbors Index**. Friends are ranked based on their number of shared neighbors with the user: $s(u, f) = ||N(u) \cap N(f)||$.
- **Inverse Friend Degree Index**. Friends with lower degrees are ranked higher i.e., $s(u,f) = \frac{1}{N(f)}$
- **Adamic-Adar Index**. This is based on the *common Neighbors* approach while penalizing each shared neighbor by its degree defined as: $s(u,f) = \sum_{a \in N(u) \cap N(f)} \frac{1}{\log|N(a)|}$
- **Resource Allocation Index**. This is based on the *Adamic-Adar* approach. It reinforces the penalization for high-degree shared neighbors. $s(u,f) = \sum_{a \in N(u) \cap N(f)} \frac{1}{|N(a)|}$
- **Preferential Attachment Index**. This models the similarity between two nodes based on their degrees and mirrors the notion of "the rich get richer". $s(u,f) = |N(u)| \times |N(f)|$
- **Jaccard Index**. This measures the ratio of common neighbors with regard to aggregate number of neighbors for two nodes. $s(u,f) = \frac{|N(u) \cap N(f)|}{|N(u) \cup N(f)|}$
- **Salton Index**. This is similar to the *Jaccard* score while the number of common neighbors is scaled with a different factor. $s(u,f) = \frac{|N(u) \cap N(f)|}{\sqrt{|N(u)| \times |N(f)|}}$
- **Sørensen Index**. This is similar to the *Jaccard* score. It reinforces the weight for friends who share higher number of neighbors with the user. $s(u,f) = \frac{2 \times |N(u) \cap N(f)|}{|N(u)| + |N(f)|}$
- **Hub Promoted Index**. This avoids to assign high score for two hub nodes solely because they share many neighbors. $s(u,f) = \frac{|N(u) \cap N(f)|}{\min(|N(u)|, |N(f)|)}$
- **Hub Depressed Index**. This promotes the ties between two hubs and between two low-degree nodes, but not between a hub and a low-degree node. $s(u,f) = \frac{|N(u) \cap N(f)|}{\max(|N(u)|, |N(f)|)}$
- **Local Leicht-Holme-Newman Index**. The tie is modeled as the ratio of the number of common neighbors and the degree of two nodes. $s(u,f) = \frac{|N(u) \cap N(f)|}{|N(u)| \times |N(f)|}$