TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 18/23 pp2237–2250
DOI: 10.26599/TST.2024.9010230
Volume 31, Number 4, August 2026

Research Article

Unifying Data Effectiveness Assessment in User Churn Detection: An Indicator-Assisted Framework

Mengying Zhou, Qianyu He, Bruno Abrahao, Xin Wang*, and Yang Chen*

Abstract: As Artificial Intelligence (AI) demonstrates an impressive performance across various tasks, its ability to process diverse data types has been widely adopted. However, AI models often lack transparency and interpretability, making it challenging to identify the most effective data categories and models for specific contexts. Current strategies typically rely on common sense or exhaustive searches, both of which are inefficient and costly. To fill this gap, we propose the INdicator-assisted FramewOrk (INFO). INFO enables researchers and practitioners to directly identify effective data categories and guide model selection for tasks involving multiple data categories. By assessing the inherent signal strength of the data, INFO provides a proxy for evaluating task performance, thereby reducing the need for time-consuming model trials and reproducibility efforts. We validate the effectiveness of INFO using real-world datasets across five different domains, focusing on the critical task of user churn detection as the case study. Our experiments show that INFO can reveal the efficacy of data categories before large-scale trials. INFO shows promise as an automated data engineering methodology that clarifies the relationship between model performance and dataset characteristics.

Key words: user churn; explainable Artificial Intelligence (AI); feature selection; data mining

1 Introduction

User churn detection is a critical task across various domains, including telecommunications[1], online gaming[2], and social media platforms[3], as it can significantly impact business growth and revenue. A 5% decrease in user churn can bring about a 25% increase in profits[4]. Accurate churn detection enables proactive interventions such as price adjustments[5], operational optimizations[2], and incentives[6].

User churn is defined as the disengagement of users from a product after a period of regular use[1, 2, 7]. Therefore, we can analyze the vast amount of data generated before user disengagement to predict churn signals. With advancements in perceptual, storage, and analysis technologies, churn signals and their informational value are being explored from diverse data types, including text, user activities, images, videos, and social connections[8, 9].

However, there is currently no unified framework to guide which data categories in a dataset are effective for churn detection, nor to specify which models should be selected for optimal performance. The sophisticated model design makes it challenging to trace the interpretability and quickly identify the most effective data categories and models in different environments. Therefore, exhaustive search among various models to find optimal solutions is the most used strategy, resulting in costly replication to assess model performance on specific datasets.

• Mengying Zhou, Qianyu He, Xin Wang, and Yang Chen are with College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433, China. E-mail: myzhou19@fudan.edu.cn; qyhe17@fudan.edu.cn; xinw@fudan.edu.cn; chenyang@fudan.edu.cn.
• Bruno Abrahao is with the Information Systems and Business Analytics, New York University, New York, NY 10012, USA. E-mail: abrahao@nyu.edu.
* To whom correspondence should be addressed.
Manuscript received: 2024-08-13; revised: 2024-11-06; accepted: 2024-11-13

More specifically, the following three factors hinder the process: **(1) Reliance on domain-specific features:** Previous approaches rely on unique domain-specific features[10, 11], which limits their generalizability across different domains. For example, some models heavily rely on linguistic data, which might not be applicable in all contexts. **(2) Varying signal strength:** Even for datasets that have the same set of features, they may encode varying levels of signal strength. For instance, social connection information might be a strong indicator for some platforms but provide weak or negligible signals on others with more impersonal interactions. **(3) Non-uniform benchmarks:** Researchers typically evaluate their approaches on specific datasets from particular domains, leading to inconsistent benchmarking across proposals. This inconsistency makes it difficult to compare approach performance across different datasets and necessitates costly replication to assess applicability in various scenarios.

This work introduces a set of interpretable indicators to unify the assessment of data effectiveness in churn detection tasks without the need for large-scale trials. These indicators allow researchers and practitioners to directly identify which data categories are most effective when applied.

The underlying design intuition behind these indicators is that the relevance of a data category to the churn detection task can be represented by its inherent informativeness. Each indicator quantifies the informativeness of its associated data category by calculating the deviation from its corresponding null model. A larger deviation from the null model indicates more informativeness. These indicators are generic and applicable across datasets and domains. The indicators are derived from five aspects: "activity periodicity", "activity diversity", "linguistic", "graph structure", and "interaction sentiment", based on data categories used in previous literature. Together, these indicators constitute the foundation of INdicator-assisted FramewOrk (INFO).

We evaluate INFO through a series of experiments using five real-world datasets from various domains. Each dataset undergoes a prior-trial assessment with INFO, followed by posterior binary classification experiments with an unbiased model. The consistency between the prior indicators and posterior results validates INFO's effectiveness. This suggests that INFO is a promising data engineering framework that clarifies the relationship between data effectiveness and informativeness. INFO can guide the selection of effective data categories and the design of sophisticated models. Our contributions are summarized as follows:

- We propose five indicators to quantify the capability of data categories in churn detection. These indicators directly map the inherent informativeness of data categories to their contribution to churn detection tasks and serve as proxies for assessing performance when applying specific data categories.
- We introduce INFO, a unified framework for data effectiveness assessment based on the above five indicators, that enables directly identifying the effective data categories without extensive model replication or testing.
- We validate INFO on five real-world datasets from diverse domains. The process involves prior-trial assessments using INFO and posterior binary classification experiments. The consistency between prior analysis and posterior results demonstrates the effectiveness of INFO.

2 Related Work

2.1 Feature selection for user churn problem

Feature selection is a crucial data preprocessing step that identifies the most useful feature sets for predictive performance. Effective models rely on robust feature selection to enhance accuracy and minimize overfitting. Zhong et al.[12] found that in essential protein prediction scenarios, more input features do not necessarily lead to better performance. By removing irrelevant or detrimental features, they achieved optimal accuracy with a subset of just eight features.

Model-based methods select features based on the model's performance. Qiu and Zhang[13] proposed a feature selection method by quantifying feature saliency according to classifier performance. In contrast, model-free methods do not rely on subsequent experimental performance. Zhou et al.[14] introduced a selection approach based on mutual information and correlation coefficient. This method calculates the correlation between candidate features and filters out those with low correlation values. Combining model-based and model-free approaches, Eljialy et al.[15] developed a hybrid feature selection technique. Their method integrated filtering techniques with wrapper methods for feature selection, which both considered feature relationships and their contributions to

experimental performance.

2.2 Modeling user churn detection

Another key relevant research topic is the construction of models with strong predictive performance for user churn. This has been studied in various fields, including the telecommunications industry[1], online games[2], and social media platforms[11, 16]. The mainstream methodology relied on feature engineering, leveraging different machine learning[2, 9] and deep learning[3] algorithms to detect possible user churn. These methods typically extract features tailored to platform-specific characteristics, such as gameplay operations in online games[2], using behavior and patterns of telecom Internet cards[1], and video preferences and topic diversity on video-sharing platforms[8].

3 Indicator Design

This section introduces the five indicators that form the foundation of INFO and their formal definitions.

3.1 Motivation

Advancements in perception, storage, and analysis technologies have enabled better task performance by utilizing multiple data categories. Using user churn detection as a case study, previous solutions have capitalized on different primary data categories based on different intuitions, as summarized in Table 1. These categories are: "activity periodicity", "activity diversity", "linguistic", "graph structure", and "interaction sentiment". Methods within the same category might employ different models but capture similar churn signals from the same data type. This presents an opportunity to measure the effectiveness of different data categories and the applicability of related models by directly comparing their inherent signal strengths. A data category with stronger signal strength contributes more significantly to task performance.

Building on this motivation, we introduce five indicators to quantify the signal strength of each data category listed in Table 1. The core idea of indicator design is to establish a null model for each category. For a given dataset, the greater the deviation of a data category from its corresponding null model, the more it contributes to task performance. The detailed definitions of these indicators are provided in the following subsections.

Table 1 Five representative data categories in churn tasks

Approach	Modeling intuition	Primary category
Wu et al.[1]	Using behavior	Activity periodicity
Yang et al.[3]	Activity pattern	
Lu et al.[8]	Video diversity	Activity diversity
Anderson et al.[17]	Musical diversity	
Lu et al.[16]	Interested topic	Linguistic
Amiri and Daumé[18]	Post content	
Óskarsdóttir et al.[9]	Social contagion	Graph structure
Piao et al.[19]	Motif network structure	
Zhang et al.[20]	Social influence	Interaction sentiment
Hamilton et al.[10]	Comment sentiment	

3.2 Indicator definitions

3.2.1 Activity periodicity

The presence of periodicity in user activity reflects the informativeness of temporal data[6]. The proposed "activity periodicity" indicator quantifies the regularity in user behavior. To establish a baseline, we use non-periodicity as the null model and consider the presence of periodicity as an indicator of valuable information for churn detection.

Datasets from different sources exhibit varying periodic granularities. For instance, Twitter activity follows a daily cycle, while data from short-term rental platforms might exhibit a monthly periodicity. To accommodate these differences, we employ a flexible and adaptive method to calculate periodicity. Specifically, we compute periodicity coefficients at daily, weekly, and monthly granularities, then select the maximum value as the final periodicity indicator,

$$\Phi = \max(\Phi^{\text{day}}, \Phi^{\text{week}}, \Phi^{\text{month}}) \tag{1}$$

where Φ^{day}, Φ^{week}, and Φ^{month} represent the periodicity coefficients for daily, weekly, and monthly granularities, respectively. These coefficients measure the average periodicity of the dataset under the corresponding time granularity and are computed as follows:

$$\Phi = \frac{\sum_{k=1}^{\left\lceil \frac{l}{g} \right\rceil - 1} \phi_p}{\left\lceil \frac{l}{g} \right\rceil - 1}, \; p = k \cdot g \tag{2}$$

Here, the granularity level g can be 1, 7, and 30, representing daily granularity, weekly granularity, and monthly granularity, respectively. The periodic correlation coefficient ϕ_p is calculated for a period

window size p. p is calculated as the product of the step k and the granularity level g. For example, with step $k=4$ and granularity level $g=7$, the period window size $p=28$ represents four weeks in terms of weekly granularity. To compute Φ, the total time span l is divided into $\left\lceil \frac{l}{g} \right\rceil$ segments, denoted as S. The correlation between these segments is measured to evaluate how well user activity conforms to periodic patterns at the given window size p. The periodic correlation coefficient ϕ_p is defined as follows:

$$\phi_p = \frac{1}{n(n-1)/2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n-1} \left| \mathrm{Corr}\left(S_{t_i:t_i+p},\, S_{t_j:t_j+p)}\right) \right| \quad (3)$$

where $n = \left\lceil \frac{l}{p} \right\rceil$. The term $\mathrm{Corr}\left(S_{t_i:t_i+p},\, S_{t_j:t_j+p)}\right)$ denotes the Pearson product-moment correlation coefficient[21] between the segments $S_{t_i:t_i+p}$ and $S_{t_j:t_j+p}$. The index t_i denotes the starting time of the i-th segment. The value of ϕ_p ranges from -1 to 1, with values closer to 1 indicating stronger periodicity.

3.2.2 Activity diversity

In addition to temporal regularity, the diversity of user behavior could also reflect the informativeness of user activity. The description of diversity spans a spectrum from generalists who explore a wide range of objects, to specialists who focus on specific areas. Here, objects and areas are broadly defined, such as specialized knowledge on Wikipedia, musical tastes on Spotify, or social interaction styles on Facebook. We apply the Generalist-Specialist score (GS-score)[22] to calculate the activity diversity for each user. The GS-score ranges from 0 to 1, where 0 represents an extreme generalist, and 1 represents an extreme specialist.

Diversity becomes informative when there is a distinct separation between generalist and specialist groups, meaning the two groups can be differentiated. Based on this principle, we define non-diversity difference as the null model and measure the differences in activity diversity distributions between the generalist and specialist groups as the indicator for the "activity diversity" category. To quantify these differences, we adopt the Wasserstein distance[23], a metric for measuring the divergence between two probability distributions. The Wasserstein distance between the generalist and specialist distributions is defined as

$$W(\mu_g, \mu_s) = \inf_{\gamma \in \Gamma(\mu_g, \mu_s)} \int c(x, y)\, \mathrm{d}\gamma(x, y) \quad (4)$$

where μ_g and μ_s represent the distributions of generalists and specialists, respectively. $\Gamma(\mu_g, \mu_s)$ denotes the set of all measures on the metric space, and $c(x, y)$ is the cost function for moving μ_g distribution to μ_s distribution. x and y are samples from μ_g and μ_s.

To obtain the diversity distributions for the generalist and specialist groups, we fit a bimodal Gaussian probability function to model the diversity distribution of all users. The two fitted Gaussian distributions μ_g and μ_s correspond to the generalist and specialist groups, respectively. The Wasserstein distance between these two Gaussian distributions quantifies the disparity between the groups and serves as the indicator for activity diversity.

3.2.3 Linguistic

Text is one of the most common data categories. Similar to activity diversity, language diversity can also provide valuable information. Language styles can vary significantly across user groups and platforms. Previous studies[10, 24] have found that churned and retained users exhibit notably different language styles. Users whose language styles are close to the community's are less likely to churn, while those with divergent styles are more likely to leave.

To quantify language style differences, we construct a Snapshot Language Model (SLM)[24]. The SLM is a bigram language model with Kneser–Ney smoothing, generated using all text posted each month. For each post po, we calculate its cross-entropy with respect to the monthly SLM $\mathrm{SLM}_{m(\mathrm{po})}$ to measure its linguistic deviation from the community, defined as follows:

$$H\left(\mathrm{po},\, \mathrm{SLM}_{m(\mathrm{po})}\right) = -\frac{1}{n_{\mathrm{po}}} \sum_{i=1}^{n_{\mathrm{po}}} \log P_{\mathrm{SLM}_{m(\mathrm{po})}}(b_i) \quad (5)$$

where $b_1, b_2, \ldots, b_n$ are the bigrams generated in post po, and $P_{\mathrm{SLM}_{m(\mathrm{po})}}(b_i)$ is the probability of each bigram b_i under this month's SLM. Lower values mean closer to the language style of the community.

For the "linguistic" data category, we define non-linguistic difference as the null model and use the linguistic differences between posts by churned and retained users as this category's indicator. This indicator measures the language style distance between the two groups, with greater differences indicating more information in this category.

To compute this indicator, we first calculate the cross-entropy for posts from churned and retained users over each time window, deriving their respective

language style distributions. We then apply the previously mentioned Wasserstein distance[23] to measure the language style differences between the two groups as indicator values.

3.2.4 Graph structure

The forms of interaction between individuals are diverse, and various graph structural features derived from these interactions are commonly used as predictors[9, 19, 25]. Graph structure is usually a result of homophily[26] and relationship closure[27]. Users in social networks tend to interact with similar people, making them more likely to behave similarly due to social influence[9].

The clustering coefficient[28] is a widely used measure of connection closeness in graph theory. It describes the likelihood of triangles to be closed, defined as the probability that any two neighbours of a given node are also connected themselves. The clustering coefficient C for a graph G is formulated as

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_G(v_i)}{\tau_G(v_i)+\lambda_G(v_i)} \tag{6}$$

where G represents the interaction graph. $\lambda_G(v_i)$ denotes the number of closed triangles on $v_i \in G$, and $\tau_G(v_i)$ is the number open triangles on $v_i \in G$. A triangle is defined as three nodes connected in pairs, while an open triangle is an incompleted triangle missing any edge.

A low clustering coefficient indicates a network close to random or less structured socially, thus less informative[29]. For the "graph structure" data category, we define random graph as the null model and use the clustering coefficient as the indicator. In a random graph, as the number of nodes N approaches $+\infty$, the clustering coefficient C_{random} tends to 0, indicating no clustering characteristic. The random graph is generated through uniform random processes, resulting in uniformly distributed information and neglective distinctions among nodes. In contrast, real networks often exhibit apparent clustering, with clustering coefficients much higher than those of random graphs of the same scale, sometimes by several orders of magnitude. This implies that real networks contain far more structured information, making the "graph structure" category a valuable data source for predictive tasks.

3.2.5 Interaction sentiment

In churn prediction tasks, sentiment is a commonly used feature[30]. Churners typically leave a platform due to negative experiences, while retained users report more positive experiences[30]. Both positive and negative interaction sentiment can provide useful churn signals, whereas neutral sentiment, with no emotional inclination, reflects a weaker signal of churn intention.

For the "interaction sentiment" data category, we define no-sentiment fluctuation as the null model and propose a biased entropy metric to measure the signal strength of interaction sentiment. Building on the classical definition of entropy, we introduce a neutral sentiment offset parameter to capture the deviation of each sentiment value from neutral sentiment. Around neutral emotions, strong sentiment polarity and divergence imply potential differentiation capability. The interaction sentiment indicator E of a given dataset is formulated as follows:

$$E = -\frac{1}{n_{\text{level}}}\sum_{i=1}^{n_s}(s_i - s_{\text{neu}})p_{s_i}\log_2 p_{s_i} \tag{7}$$

where s_i denotes the i-th sentiment value, and p_{s_i} is the probability of occurrence of sentiment value s_i. n_s is the number of distinct sentiment values. s_{neu} denotes the absolute neutral sentiment value, typically calculated as $\frac{s_{\max}+s_{\min}}{2}$. In this work, sentiment scores range from minimum value $s_{\min}=0$ to maximum value $s_{\max}=1$, and the absolute neutral value of s_{neu} is 0.5. $j - s_{\text{neu}}$ is the neutral sentiment offset parameter. n_{level} is specified as 5, representing the number of sentiment levels[31].

4 Prior-Trial Assessment with INFO

Using the indicators proposed above, we develop the INFO to assess the effectiveness and contribution of various data categories in a dataset for the churn prediction task. In this section, we apply INFO to perform prior-trial assessments on user churn detection using five real-world datasets. The results demonstrate that INFO can effectively reflect the informativeness in different data categories.

4.1 Datasets description

To ensure the comprehensiveness and generalizability of the evaluation, we use a diverse collection of five datasets from various domains. The diversity of the target datasets is ensured based on the following two criteria. **(1) Diversity in functional categories.** This refers to datasets originating from different platform types. We consider platforms including online lodging, code hosting, online encyclopedias, and location-based

services. Each platform provides unique services with distinct user behavior patterns and data categories. **(2) Diversity in functional contexts.** This accounts for different inherent motivations driving user interactions. Platforms offering similar services exhibit different functional characteristics, such as transaction-based hospitality on Airbnb and non-monetary host-guest exchanges on Couchsurfing[32]. This diversity allows us to discuss how the efficacy of the same data category varies across different contexts.

- **Airbnb** is an online lodging platform where hosts list rooms with prices, and guests reserve places. It facilitates trade and socialization through short-term rentals. Hosts and guests can review each other after check-out. We collected transaction and review data from Airbnb website from January 1, 2019, to March 19, 2019, via breadth-first search.
- **Couchsurfing** is a non-profit travel community where users share spare rooms for free, emphasizing social interaction. Hosts and guests exchange reviews and ratings after stays. Data is sourced from Ref. [33].
- **GitHub** is a code-hosting community that promotes collaboration among developers. Users manage files using the git tool, which tracks activities such as commits, pushes, and merges. Users can also follow other developers and projects to track updates. Data is available from Refs. [34, 35].
- **Wikipedia** is the largest multilingual user-contributed online encyclopedia. Both registered and anonymous users can edit entries and engage in discussions. Long-term editors use watchlists to monitor changes to specific entries. Data is available from Ref. [36].
- **Foursquare** is a location-based service platform where users post check-ins at venues[37]. They can rate venues, provide reference information for others, and add friends. Data is available from Ref. [38].

Table 2 summarizes the statistics of the five diverse datasets. Each dataset covers over 27 000 users, with activity records reaching up to one million. For Airbnb, Couchsurfing, GitHub, and Wikipedia, we use churn prediction window sizes of 6 or 12 months. For Foursquare, a one-month prediction window is used due to its shorter time span of only five months.

Table 2 Statistic information of datasets.

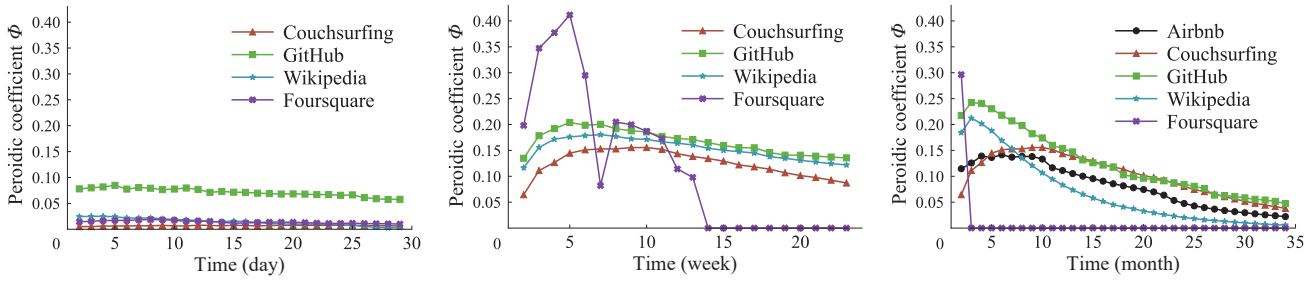
Dataset	Number of users	Number of records	Time span (month)	Prediction window size (month)
Airbnb	29 138	708 020	130	12
Couchsurfing	49 167	157 822	70	6
GitHub	122 239	8 407 824	124	12
Wikipedia	32 819	1 835 055	82	12
Foursquare	27 304	184 723	5	1

4.2 Assessment results

4.2.1 Periodicity at different granularities

For activity periodicity, we plot the periodicity correlation coefficients ϕ at daily, weekly, and monthly granularities g in Fig. 1. Higher values on the y-axis indicate stronger periodicity for the corresponding window size. Note that for Airbnb, only monthly granularity is analyzed due to its transaction records being available exclusively at the monthly level.

The results reveal that periodicity varies across datasets and granularities. Both Airbnb and Couchsurfing exhibit apparent periodicity at the monthly granularity, reflecting user behaviors tied to travel patterns. GitHub's periodicity is observed across all granularities, likely due to its role as a code collaboration platform where developers exhibit strong daily activity patterns[39]. Wikipedia displays apparent periodicity at weekly and monthly granularities, highlighting patterns in editing activities. Foursquare exhibits periodicity only at the weekly granularity, with significant fluctuations. This could be related to the dataset's shorter time span of just five months, limiting



(a) Daily granularity (g = 1) (b) Weekly granularity (g = 7) (c)Monthly granularity (g = 30)

Fig. 1 Periodicity coefficient Φ at different time granularities.

its ability to capture long-term user behavior.

More importantly, INFO can identify the most effective periodic window size at a finer granularity. Each dataset has a peak representing the strongest periodicity. Airbnb and Couchsurfing achieve their peaks at around 10 months, consistent with annual travel seasonality patterns[40]. GitHub and Wikipedia show peaks in three months, reflecting quarterly engagement activities like maintenance and updates. These results highlight INFO's capability to guide the selection of appropriate periodic granularity for downstream tasks. Choosing an inappropriate granularity may fail to capture valuable information about periodicity.

4.2.2 Diversity distribution patterns

As defined in Section 3.2.2, the diversity indicator assesses the difference in diversity distribution between generalist and specialist groups. Figure 2 shows the diversity distributions of the five datasets, along with their fitted bimodal Gaussian functions. It is notable that not all datasets exhibit a bimodal structure: Couchsurfing and GitHub display two clear peaks representing generalists and specialists, while Wikipedia's bimodal are less pronounced. In contrast, Airbnb and Foursquare exhibit only one visible peak, indicating a lack of differentiation among users.

Particularly, the comparison between Couchsurfing and Airbnb highlights the diversity indicator's ability to reflect different churn signal strengths. On Couchsurfing, both groups have diversity scores below 0.5, reflecting their active exploration of diverse interactions due to its non-monetary rental model. In contrast, Airbnb, as a preferred accommodation website, has gradually transitioned from a sharing platform to a more specialized accommodation service[41], exhibiting lower user diversity scores clustering around a single peak.

Unlike Couchsurfing and Airbnb, GitHub and Wikipedia are typical knowledge-based expert platforms, where most users focus on specific areas of interest. Therefore, compared with exploratory platforms, their diversity score distributions are close to 1, indicating that their users are more like experts. GitHub still exhibits bimodal distributions due to the presence of users mastering multiple programming languages. Foursquare, being a check-in platform, shows high regularity in user trajectories, and check-in locations are relatively concentrated, resulting in diversity scores close to 1.

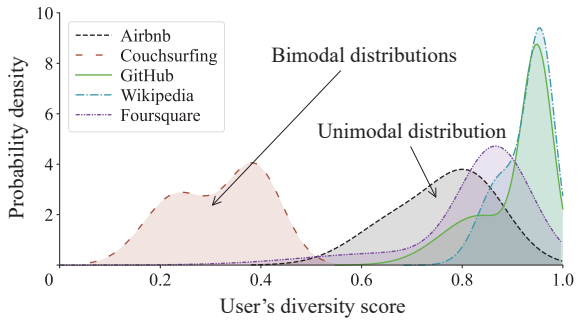


Fig. 2 Bimodal distributions of each dataset.

4.2.3 Linguistic distance

We analyze the language style difference between churned and retained users as a linguistic indicator. The intuition is that potential churned users will gradually deviate from the community's language style. The language styles are represented by the average cross-entropy, calculated based on the monthly SLM and posted text.

Figure 3 shows the cross-entropy of the language style for churned and retained users on Airbnb and GitHub. On Airbnb, the difference between the two groups is small, while on GitHub, the difference is more significant and widens over time. Table 3 quantifies these differences, with Airbnb's linguistic indicator value at 0.018 and GitHub's at 0.075. This indicates that churned and retained users on Airbnb have similar language styles, whereas churned users on GitHub exhibit significantly different language styles, making linguistic data more useful for churn detection on GitHub.

An interesting observation is the cyclic variation in language cross-entropy on Airbnb, in contrast to the monotonic increase on GitHub. This reflects annual travel patterns on Airbnb. During peak travel months (July and August), the influx of seasonal users diversifies language styles. As these users leave,

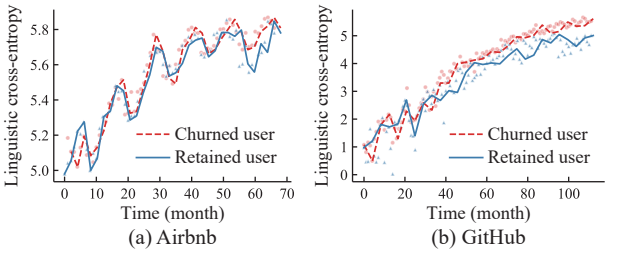


Fig. 3 Cross-entropy of the language style over time for churned and retained users on Airbnb andGitHub. The scattered points represent individual monthly data, while the lines show the smoothed trends fitted from those points.

Table 3 Indicator values and churn detection performance of the five datasets. r_{intra} and r_{inter} are the Pearson correlation coefficients between the indicators and the performance within and across datasets, respectively. Some datasets lack certain feature types, and their values are denoted as N/A.

Data category	Airbnb		Couchsurfing		GitHub		Wikipedia		Foursquare		r_{inter}
	Indicator	F1-score	Indicator	F1-score	Indicator	F1-score	Indicator	F1-score	Indicator	F1-score	
Activity periodicity	0.108	0.704	0.137	0.756	0.197	0.784	0.149	0.768	0.112	0.728	0.909
Activity diversity	0.035	0.680	0.065	0.732	0.053	0.726	0.028	0.619	0.053	0.711	0.924
Linguistic distance	0.018	0.670	N/A	N/A	0.075	0.689	0.036	0.066	N/A	N/A	0.767
Graph structure	0.001	0.665	0.133	0.751	0.009	0.686	N/A	N/A	0.105	0.725	0.519
Interaction sentiment	0.136	0.748	0.076	0.742	0.049	0.714	0.036	0.648	0.072	0.727	0.755
r_{intra}	0.953		0.831		0.890		0.634		0.854		N/A

language styles stabilize and become more uniform.

4.2.4 Clustering level of graph structure

The graph construction methods for each dataset are as follows. The Wikipedia dataset cannot construct a graph due to the absence of relationship data.

- **Airbnb: Transactions graph.** Nodes represent users and undirected edges represent the transactions between users.
- **Couchsurfing: Interaction graph.** Nodes represent users and undirected edges represent the non-profit lodging interactions between users.
- **GitHub: Follower-followee graph.** Nodes represent users and directed edges represent the follower-followee relationships.
- **Foursquare: Friendship graph.** Nodes represent users and undirected edges represent the friendship connections between users.

The clustering coefficients for each dataset are listed in Table 3. Significant differences are observed in the clustering coefficients across the four datasets. The Airbnb graph has the lowest clustering coefficient at 0.001, while the Couchsurfing graph, a similar short-rental platform, has a higher coefficient of 0.133. This disparity is attributed to the difference between transactional interactions on Airbnb and friendship-based interactions on Couchsurfing. GitHub and Foursquare have clustering coefficients of 0.009 and 0.105, respectively. These findings suggest that graphs based on "friendship" or "interest" interactions tend to be denser and more informative compared with those based on transactions. This assumption is further supported by the experimental results in Section 5.

4.2.5 Interaction sentiment distribution

In our five datasets, interaction sentiment is derived from numerical and textual data. Couchsurfing and Foursquare inherently include numerical rating values. For Airbnb, GitHub, and Wikipedia, we use VADER[42], a sentiment analysis tool for social media text, to compute sentiment scores ranging from 0 (most negative) to 1 (most positive) based on the textual data in interactions or activities.

Table 3 lists the interaction sentiment indicators for all datasets. Collaborative platforms like GitHub and Wikipedia show lower indicator values, as interactions are primarily objective and knowledge-focused, with smaller deviations from neutral sentiment. In contrast, entertainment and social platforms like Airbnb, Couchsurfing, and Foursquare exhibit higher indicator values, reflecting stronger emotional expressions and more informativeness.

Figure 4 further illustrates each dataset's Cumulative Distribution Function (CDF) of sentiment, visually showing deviations from neutrality. Sentiment values on GitHub and Wikipedia concentrate around 0.5, reflecting their neutral, objective, knowledge-centered interactions. Meanwhile, sentiment on Airbnb, Couchsurfing, and Foursquare shows stronger emotions with three different patterns: dominant-positive for Airbnb, dominant-negative for Couchsurfing, and relatively evenly distributed for Foursquare.

5 Posterior Experiment

In the previous section, we have extensively analyzed the efficacy of various data categories using the five indicators of INFO. In this section, we perform a binary classification on user churn prediction and compare the experimental results with the prior-trial

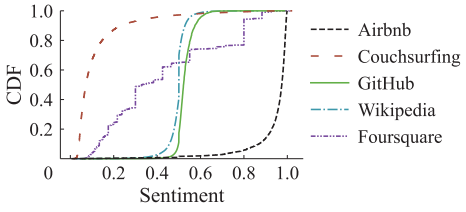


Fig. 4 CDF of interaction sentiment.

indicators to answer the following research questions:

- **RQ1:** Is there a clear consistency between the prior-trial indicators provided by INFO and the posterior experimental results?
- **RQ2:** Can the prior-trial indicators from INFO assist in guiding model selection and feature engineering?
- **RQ3:** Can INFO directly quantify the performance of a specific model using a particular data category across different datasets?

5.1 Experimental settings

INFO is designed to reveal the inherent churn signal in raw data. To minimize bias from complex feature engineering, we extract commonly used tabular features. For instance, in the graph structure feature category, we use ego networks instead of the entire graph to minimize complexity. The detailed features used are listed in Table 4. Given that churn detection is a binary classification task, we use a non-parametric Logistic Regression (LR) classifier, which also achieves model simplicity and avoids parameter tuning bias. We evaluate performance using the classic metric F1-score.

5.2 Prior indicator vs posterior experiment

Each feature category is independently used in the LR model, with its performance reflecting the effectiveness of the corresponding data category. Table 3 summarizes the experiment results alongside the prior-trial indicator values obtained using INFO.

High consistency between prior indicators and posterior performance. The Pearson correlation coefficient is used to measure the correlation between prior indicators and posterior experiment performance within each dataset. These results, denoted as r_{intra}, are shown in the last row of Table 3. Most values of r_{intra} exceed 0.8, indicating a strong positive correlation between prior indicators and posterior experiment performance. This high consistency underscores INFO's ability to reveal the potential of feature categories without replicating models reliant on those categories.

Additionally, we note that the most effective data categories vary across datasets. For Airbnb, interaction sentiment has the strongest signal, while for the other four datasets, activity periodicity is the dominant factor. This suggests that user churn detection tasks on different platforms rely on different data categories. It also explains why some well-trained models may not be applicable across all scenarios, emphasizing the importance of choosing an appropriate model tailored to the characteristics of each dataset.

Ability to capture periodicity at different granularities. As discussed in Section 4.2.1, INFO shows that different datasets prefer different periodic granularities. Here, we verify INFO's ability to adaptively capture periodicity across granularities and examine the impact of selecting appropriate granularity on prediction performance.

Table 5 presents the periodicity indicators at three granularities and their corresponding experimental performance. The optimal granularity varies by dataset. For Airbnb, Couchsurfing, and GitHub, monthly granularity yields the best indicators and performance. Conversely, for Wikipedia and Foursquare, weekly granularity is most effective. These findings align with the different user activity patterns on each platform.

Interestingly, using coarser granularities can achieve comparable or even superior classification performance compared with finer granularities. On Couchsurfing and GitHub, the F1-scores for monthly granularity exceed those for weekly granularity, suggesting that monthly features are sufficient to provide helpful information. Additionally, performance at weekly

Table 4 Specific features used in the experiments.

Data category	Feature name	Description
Activity periodicity	Activity record	Number of records at each time slot
Activity diversity	GS-score	GS-score value
Linguistic	Cross-entropy	Average cross-entropy
Graph structure	Size	Number of nodes that one-step out neighbors of ego node
	Ties	Number of connections with all the nodes in the ego network
	Pairs	Number of possible directed ties in the ego network
	Density	Number of ties divided by number of pairs
Interaction sentiment	Sentiment	Sentiment value

Table 5 Impact of time granularity quality on performance. We respectively list the periodicity indicators of the five datasets at the daily, weekly, and monthly granularities and the corresponding prediction performance. Some datasets lack certain feature types, and their values are denoted as N/A.

Granularity	Airbnb		Couchsurfing		GitHub		Wikipedia		Foursquare	
	Φ	F1-score	Φ	F1-score	Φ	F1-score	Φ	F1-score	Φ	F1-score
Daily	N/A	N/A	0.014	0.748	0.124	0.776	0.012	0.736	0.017	0.727
Weekly	N/A	N/A	0.113	0.752	0.165	0.777	0.149	0.768	0.112	0.725
Monthly	0.108	0.730	0.137	0.756	0.197	0.784	0.143	0.766	0.099	0.722

granularity often outperforms that at daily granularity, indicating that finer granularities are not always necessary. Appropriately coarse-grained data already can effectively capture user behavior periodicity while reducing computational resource demands.

Prioritize promising features. Beyond identifying the most effective data categories, INFO also helps prioritize promising features, offering guidance for optimizing model performance.

Each feature of the five data categories is assigned a weight based on its indicator value. Using these weights, we perform feature selection through a weight-based method and compare it to random selection. We keep the same percentage of features as model input for both methods. The performance curves for the two selection methods are shown in Fig. 5.

Features selected based on indicator weights consistently outperform those selected randomly. Notably, on GitHub and Foursquare, even a small percentage of weight-based selected features can achieve competitive performance compared with using all features. On Airbnb, selecting just 40% of the features yields the best performance. These results highlight INFO's ability to assist researchers and practitioners in prioritizing promising features. This is particularly valuable in scenarios where acquiring all features is challenging, as it reduces trial-and-error costs in practical scenarios.

5.3 Quantifying naive model performance across different datasets

The above results demonstrate INFO's effectiveness in identifying the potential of different data categories within a dataset. In this subsection, we explore another question: can INFO directly quantify the performance of a specific model across different datasets? The intuition is that a model relying on a particular data category should perform well on datasets where that category has strong indicator values. In other words, when the model is built on a specific data category, and the corresponding indicator for that data category on the given dataset shows strong signals, the model should be transferable to that dataset or domain.

Previous churn detection methods often leverage multiple data categories and involve sophisticated model designs[3, 10]. Due to INFO's focus on comparing the naive model based on a single data category, our analysis is limited to comparing the performance of naive models. For this purpose, we employ an unbiased LR model built on a single data category as the naive model, as described in the experimental setup.

In Table 3, the rightmost column reports the Pearson correlation coefficients r_{inter} between indicator values and the performance of naive models across datasets. These results derive the following findings.

Positive correlation between indicators and naive model performance across datasets. The listed

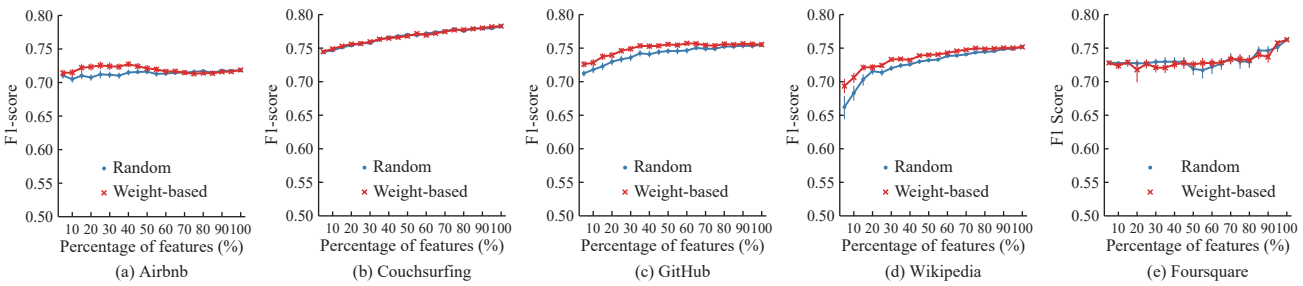


Fig. 5 Performance in F1-score with the selected feature percentage. We use random selection and weight-based selection according to indicators to select a certain percentage of features for prediction in our extracted feature set.

Pearson correlation coefficients r_{inter} are mostly above 0.7, indicating strong alignment between the indicators and experimental results across datasets. However, the graph structure category shows a lower r_{inter} value of 0.519. This discrepancy might be attributed to the unique characteristics of graph-based data. The effectiveness of graph features depends heavily on the nature and definition of connections, which vary significantly across datasets. Such variability reduces consistency when applying naive graph-based models across datasets.

Friendship makes graph structure informative. Social networking is a core functionality of many online platforms. Graph structure models perform better on datasets where connections are based on friendships. For instance, graph-based models perform poorly on Airbnb but perform well on Couchsurfing. This difference arises from the nature of interactions: Airbnb is a monetary-based platform focused on service quality, whereas Couchsurfing emphasizes friend-based relationships between guests and hosts[32]. Naive models applied to other datasets also show that friendship-based graphs outperform those based on transactional interactions. These findings validate the role of social contagion in user churn and highlight that monetary interactions do not equate to friendship connections.

Churn signals displayed in emotion. Sentiment is a crucial measure of interaction quality. Platforms that emphasize user experiences (e.g., Airbnb) or friendships (e.g., Couchsurfing and Foursquare) exhibit better prediction performance than those centered on objective interactions (e.g., GitHub and Wikipedia). This is consistent with previous research considering sentiment as a strong predictor of churn[10, 20]. Notably, both text-based and rating-based sentiment data show high predictive ability. Textual sentiment on Airbnb is as effective as, and sometimes more effective than, ratings. Text can capture the implicit negative sentiment that ratings may fail to reflect, complementing missing sentiment information.

6 Discussion

6.1 INFO's performance on incomplete datasets with missing columns

In the real world, datasets often suffer from the "incomplete feature sets" problem due to factors such as data collection limitations, privacy concerns, or operational constraints. Missing indicators will result in incomplete evaluations of the informativeness of data categories, potentially weakening the INFO's predictive capability. Ensuring INFO's robustness in handling incomplete datasets is essential for its practical applicability in real-world scenarios. This subsection explores strategies to address challenges posed by datasets with missing columns.

When datasets lack certain features, INFO can infer missing signals from the available data. For instance, in domains with sparse user interactions, such as financial transactions or sensor-based applications, alternative features derived from non-interactive data can reveal relevant patterns. In financial transactions, features such as transaction frequency, volume, or temporal trends can replace or supplement interaction-based features. These alternatives provide insights into user behavior, enabling INFO to detect churn signals through shifts in activity patterns rather than direct interactions. By leveraging existing signals, INFO maintains its assessment integrity and extends its applicability to churn detection across a wider range of domains.

6.2 Applicability of INFO in broader applications

In this work, INFO has demonstrated effectiveness in user churn detection. Moreover, its potential can be extended to wider applications. Its interpretable indicators can provide valuable insights in tasks like customer segmentation, fraud detection, and predictive maintenance as well.

Customized indicators combinations for broader predictive tasks. INFO's indicators can be tailored for various predictive tasks beyond churn detection. For example, in recommendation systems, "activity periodicity" and "activity diversity" can be used in user preference prediction tasks to understand the contribution to recommendation accuracy. In fraud detection, irregularities in "graph structure" and unexpected "activity patterns" might signal fraudulent activities. In market segmentation, "linguistic" and "interaction sentiment" indicators can provide insights into communication styles and emotional engagement, aiding in customer group identification and targeted marketing strategies. By adapting INFO's indicators to specific tasks, INFO can offer valuable insights and improve performance across various applications.

Adjusting indicator weightings. Additionally,

INFO allows for flexible adjustment of indicator weightings to meet the requirements of specific predictive tasks. By prioritizing the most relevant indicators for the given application, INFO can be fine-tuned to provide targeted guidance. For instance, in fraud detection, increasing the weight of indicators related to anomalous transactions or irregular network structures can enhance their relevance for identifying fraudulent behaviors. By extending and adjusting INFO's indicators and their weightings, INFO can be more versatile and broadly applicable.

7 Conclusion and Future Work

This paper introduces INFO, a unified framework to directly assess and select effective data categories for specific application domains. INFO quantifies the relevance and effectiveness of data categories in churn detection tasks by measuring their inherent informativeness. We identify five data categories and formulate their corresponding indicators as the foundation of INFO. Through experiments on five datasets from diverse domains, we demonstrate INFO's capability to directly quantify the prediction performance of data categories. The strong consistency between prior indicators and posterior experiment results validates INFO's effectiveness and reliability.

Our future work will extend INFO to hybrid data categories to enhance its generalization, evaluate its robustness across broader scenarios, and explore general principles of user stickiness and retention.

Acknowledgment

This work was sponsored by the National Natural Science Foundation of China (Nos. 62072115 and 62472101), Shanghai Science and Technology Innovation Action Plan Project (No. 22510713600), Shanghai Municipal Science and Technology Commission (No. 22dz1204900), the Fundamental Research Funds for the Central Universities (No. 2025110481), and Shanghai Key Laboratory of Intelligent Information Processing, Fudan University (No. IIPL-2025-RD1-03).

References

[1] F. Wu, F. Lyu, J. Ren, P. Yang, K. Qian, S. Gao, and Y. Zhang, Characterizing internet card user portraits for efficient churn prediction model design, *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1735–1752, 2024.

[2] Y. Xiong, R. Wu, S. Zhao, J. Tao, X. Shen, T. Lyu, C. Fan, and P. Cui, A data-driven decision support framework for player churn analysis in online games, in *Proc. 29th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, New York, NY, USA, 2023, pp. 5303–5314.

[3] C. Yang, X. Shi, L. Jie, and J. Han, I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application, in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, New York, NY, USA, 2018, pp. 914–922.

[4] F. Reichheld and C. Detrick, Loyalty: A prescription for cutting costs, *Mark. Manag.*, vol. 12, no. 5, pp. 24–25, 2003.

[5] P. Ye, J. Qian, J. Chen, C. Wu, Y. Zhou, S. De Mars, F. Yang, and L. Zhang, Customized regression model for Airbnb dynamic pricing, in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 932–940.

[6] R. Jakob, N. Lepper, E. Fleisch, and T. Kowatsch, Predicting early user churn in a public digital weight loss intervention, in *Proc. of the CHI Conf. Human Factors in Computing Systems*, Honolulu, HI, USA, 2024, p. 994.

[7] H. Sagtani, M. G. Jhawar, A. Gupta, and R. Mehrotra, Quantifying and leveraging user fatigue for interventions in recommender systems, in *Proc. 46th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Taipei, China, 2023, pp. 2293–2297.

[8] Y. Lu, P. Cui, L. Yu, L. Li, and W. Zhu, Uncovering the heterogeneous effects of preference diversity on user activeness: A dynamic mixture model, in *Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Washington, DC, USA, 2022, pp. 3458–3467.

[9] M. Óskarsdóttir, K. E. Gísladóttir, R. Stefánsson, D. Aleman, and C. Sarraute, Social networks for enhanced player churn prediction in mobile free-to-play games, *Appl. Netw. Sci.*, vol. 7, no. 1, p. 82, 2022.

[10] W. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec, Loyalty in online communities, in *Proc. 11th Int. AAAI Conf. Web and Social Media*, Montreal, Canada, 2017, pp. 540–543.

[11] T. Althoff and J. Leskovec, Donor retention in online crowdfunding communities: A case study of DonorsChoose.org, in *Proc. 24th Int. Conf. World Wide Web*, Florence, Italy, 2015, pp. 34–44.

[12] J. Zhong, J. Wang, W. Peng, Z. Zhang, and M. Li, A feature selection method for prediction essential protein, *Tsinghua Science and Technology*, vol. 20, no. 5, pp. 491–499, 2015.

[13] Y. Qiu and C. Zhang, Research of indicator system in customer churn prediction for telecom industry, in *Proc. 11th Int. Conf. Computer Science & Education*, 2016, pp. 123–130.

[14] H. Zhou, X. Wang, and R. Zhu, Feature selection based on mutual information with correlation coefficient, *Appl. Intell.*, vol. 52, no. 5, pp. 5457–5474, 2022.

[15] A. E. M. Eljialy, M. Y. Uddin, and S. Ahmad, Novel framework for an intrusion detection system using multiple feature selection methods based on deep learning, *Tsinghua Science and Technology*, vol. 29, no. 4, pp. 948–958, 2024.

[16] Y. Lu, L. Yu, P. Cui, C. Zang, R. Xu, Y. Liu, L. Li, and W. Zhu, Uncovering the co-driven mechanism of social and content links in user churn phenomena, in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 3093–3101.

[17] A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas, Algorithmic effects on the diversity of consumption on Spotify, in *Proc. Web Conf.*, Taipei, China, 2020, pp. 2155–2165.

[18] H. Amiri and H. Daumé III, Short text representation for detecting churn in microblogs, in *Proc. 30th AAAI Conf. Artificial Intelligence*, Phoenix, AZ, USA, 2016, pp. 2566–2572.

[19] J. Piao, G. Zhang, F. Xu, Z. Chen, and Y. Li, Predicting customer value with social relationships via motif-based graph attention networks, in *Proc. of the Web Conf.*, Ljubljana, Slovenia, 2021, pp. 3146–3157.

[20] G. Zhang, J. Zeng, Z. Zhao, D. Jin, and Y. Li, A counterfactual modeling framework for churn prediction, in *Proc. 15th ACM Int. Conf. Web Search and Data Mining*, Virtual Event, 2022, pp. 1424–1432.

[21] K. Pearson, Note on regression and inheritance in the case of two parents, *Proc. Roy. Soc. London*, vol. 58, no. 347-352, pp. 240–242, 1895.

[22] I. Waller and A. Anderson, Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 1954–1964.

[23] L. V. Kantorovich, Mathematical methods of organizing and planning production, *Manag. Sci.*, vol. 6, no. 4, pp. 366–422, 1960.

[24] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, No country for old members: User lifecycle and linguistic change in online communities, in *Proc. 22nd Int. Conf. World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 307–318.

[25] F. Qian, Y. Gao, S. Zhao, J. Tang, and Y. Zhang, Combining topological properties and strong ties for link prediction, *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 595–608, 2017.

[26] B. Abrahao, P. Parigi, A. Gupta, and K. S. Cook, Reputation offsets trust judgments based on social biases among Airbnb users, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 37, pp. 9848–9853, 2017.

[27] G. Kossinets and D. J. Watts, Empirical analysis of an evolving social network, *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[28] J. Saramäki, M. Kivelä, J. P. Onnela, K. Kaski, and J. Kertész, Generalizations of the clustering coefficient to weighted complex networks, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 75, no. 2, p. 027105, 2007.

[29] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, Microscopic evolution of social networks, in *Proc. 14th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Las Vegas, NV, USA, 2008, pp. 462–470.

[30] X. Wang, J. Zhang, C. Gu, and F. Zhen, Examining antecedents and consequences of tourist satisfaction: A structural modeling approach, *Tsinghua Science and Technology*, vol. 14, no. 3, pp. 397–406, 2009.

[31] E. Guzman and W. Maalej, How do users like this feature? A fine grained sentiment analysis of app reviews, in *Proc. 22nd Int. Requirements Engineering Conf.*, Karlskrona, Sweden, 2014, pp. 153–162.

[32] M. Klein, J. Zhao, J. Ni, I. Johnson, B. M. Hill, and H. Zhu, Quality standards, service orientation, and power in Airbnb and Couchsurfing, in *Proc. ACM on Human-Computer Interaction, Volume 1, Issue CSCW*, Virtual Event, 2017, p. 58.

[33] B. State, B. Abrahao, and K. Cook, Power imbalance and rating systems, in *Proc. 10th Int. AAAI Conf. Web and Social Media*, Cologne, Germany, 2016, pp. 368–377.

[34] Q. Gong, Y. Liu, J. Zhang, Y. Chen, Q. Li, Y. Xiao, X. Wang, and P. Hui, Detecting malicious accounts in online developer communities using deep learning, *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10633–10649, 2023.

[35] Q. Gong, J. Zhang, Y. Chen, Y. Xiao, X. Fu, P. Hui, X. Li, and X. Wang, A representative user-centric dataset of 10 million GitHub developers, https://research.aalto.fi/en/datasets/a-representative-user-centric-dataset-of-10-million-github-develo, 2018.

[36] J. Leskovec. Complete Wikipedia edit history (up to January 2008), http://snap.stanford.edu/data/wiki-meta.html, 2008.

[37] Y. Chen, J. Hu, Y. Xiao, X. Li, and P. Hui, Understanding the user behavior of Foursquare: A data-driven study on a global scale, *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 4, pp. 1019–1032, 2020.

[38] Sarwat foursquare dataset, https://archive.org/details/201309_foursquare_dataset_umn, 2013.

[39] J. Zhang, Y. Chen, Q. Gong, X. Wang, A. Y. Ding, Y. Xiao, and P. Hui, Understanding the working time of developers in IT companies in China and the United States, *IEEE Softw.*, vol. 38, no. 2, pp. 96–106, 2021.

[40] Q. Zhou, Y. Chen, C. Ma, F. Li, Y. Xiao, X. Wang, and X. Fu, Measurement and analysis of the reviews in Airbnb, in *Proc. IFIP Networking Conf.* (*IFIP Networking*) *and Workshops*, Zurich, Switzerland, 2018, pp. 1–9.

[41] G. Quattrone, D. Proserpio, D. Quercia, L. Capra, and M. Musolesi, Who benefits from the "sharing" Economy of Airbnb? in *Proc. 25th Int. Conf. World Wide Web*, Montreal, Canada, 2016, pp. 1385–1394.

[42] C. Hutto and E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in *Proc. 8th Int. AAAI Conf. Web and Social Media*, Ann Arbor, MI, USA, 2014, pp. 216–225.

Mengying Zhou received the BS degree in information security from Lanzhou University, China, in 2019, and received the PhD degree in computer science from Fudan University, China, in 2025. She is an assistant professor at School of Computing and Artificial Intelligence, Shanghai University of Finance and Economics, China. Her research interests include edge MLSys and urban computing.

Bruno Abrahao received the PhD degree in computer science from Cornell University, Ithaca, NY, USA, in 2014. He is currently an assistant professor at New York University, New York, NY, USA. He is also a faculty member at the Center for Artificial Intelligence and Deep Learning, New York University Shanghai, Shanghai, China. He was a postdoctoral researcher at Stanford University, Stanford, CA, USA, and worked at Microsoft Research AI, Redmond, WA, USA. His research interests include information systems, business analytics, and global network.

Xin Wang received the BS degree in information theory and the MS degree in communication and electronic systems from Xidian University, China, in 1994 and 1997, respectively, and received the PhD degree in computer science from Shizuoka University, Japan, in 2002. He is currently a professor at College of Computer Science and Artificial Intelligence, Fudan University, China. His main research interests include quality of network service, next generation network architecture, mobile Internet and network coding.

Qianyu He received the BS degree from Fudan University, China, in 2021. She is currently a PhD candidate at Fudan University, China. Her primary research focus on instruction following and mathematical reasoning ability of LLMs to improve their practicality and reliability.

Yang Chen received the BS and PhD degrees from Tsinghua University, China, in 2004 and 2009, respectively. He is currently a professor at College of Computer Science and Artificial Intelligence, Fudan University, China, and a vice director of the Shanghai Key Lab of Intelligent Information Processing, China. He was a postdoctoral researcher at Department of Computer Science, Duke University, Durham, NC, USA. He is serving as an associate editors-in-chief of *Journal of Social Computing*, a senior associate editor of *ACM Transactions on Social Computing*, and an associate editor of *Computer Communications*. He served as an OC/TPC member for many international conferences, including SOSP, SIGCOMM, WWW, MobiSys, ICDCS, IJCAI, AAAI, IWQoS, ICCCN, GLOBECOM and ICC. He published more than 50 referred papers in international journals and conferences, including TPDS, TDSC, TMC, TKDE, TSC, TCSS, TNSM, IEEE Communications Magazine, IEEE Software, Middleware, INFOCOM, ICDE, COSN, CIKM, and IWQoS. His research interests include social computing, Internet architectures and mobile computing.