Analysis of user popularity pattern and engagement prediction in online social networks

Samin Mohammadi

Doctor of Philosophy (PhD) Thesis

Sorbonne Université and Télécom SudParis

Specialization

Computer Science - Artificial Intelligence

National Thesis Number (NNT)

2018TELE0019

presented by

Samin Mohammadi

Analysis of User Popularity Pattern and Engagement Prediction in Online Social Networks

4^{th} December 2018

Committee:

Markus Fiedler	Reviewer	Professor, Blekinge Tekniska Hogskola University - Sweden
Agata Filipowska	Reviewer	Assistant Professor, Poznan University of Economics - Poland
Ioan Marius Bilasco	Examiner	Associate Professor, Université Lille 1, France
Marie-Jeanne Lesot	Examiner	Associate Professor, Sorbonne Université - France
Yacine Ghamri-Doudane	Examiner	Professor, Université de La Rochelle - France
Daqing Zhang	Examiner	Professor, Institut Télécom SudParis - France
Patrick Constant	Examiner	CEO of Pertimm, Pertimm - France
Noel Crespi	Supervisor	Professor, Institut Télécom SudParis - France
Reza Farahbakhsh	Co-supervisor	Research Scientist, Institut Télécom SudParis - France

Thèse de Doctorat (PhD) de

Sorbonne Université et Institut Télécom SudParis

Spécialité

Informatique

Numéro National de Thése (NNT)

2018TELE0019

présentée par

Samin MOHAMMADI

Analyse du modéle de popularité de l'utilisateur et de la prédiction d'engagement en les réseaux sociaux en ligne

4^{th} Decembre 2018

Jury composé de :

Markus Fiedler	Rapporteur	Professeur, Blekinge Tekniska Hogskola University - Sweden
Agata Filipowska	Rapporteur	Professeur Assistant, Poznan University of Economics - Poland
Ioan Marius Bilasco	Examinateur	Maitre de Conférence, Université Lille 1, France
Marie-Jeanne Lesot	Examinateur	Maitre de Conférence, Sorbonne Université- France
Yacine Ghamri-Doudane	Examinateur	Professeur, Université de La Rochelle - France
Daqing Zhang	Examinateur	Professeur, Institut Télécom SudParis - France
Patrick Constant	Examinateur	Président directeur général de Pertimm, Pertimm - France
Noel Crespi	Directeur	Professeur, Institut Télécom SudParis - France
Reza Farahbakhsh	Encadrant	Chercheur Scientifique, Institut Télécom SudParis - France

Acknowledgements

I would like to express my sincere thanks to Professor Noel CRESPI, my thesis supervisor at Institut Télécom SudParis, who guided my researches and answered my questions. I would like to thank him for all his help, it was absolutely invaluable.

More thanks go to my thesis reviewers for all their invaluable comments and guidance, Professor Markus Fiedler and Professor Agata Filipowska.

I would especially like to thank my thesis committee members composed of Dr. Ioan Marius BILASCO, Dr. Marie-Jeanne LESOT, Professor Daqing ZHANG, Professor Yacine GHAMRI-DOUDANE, and Dr. Patrick CONSTANT.

I would also like to thank my co-supervisor, all my friends and colleagues who helped me during my PhD.

Above all, many thanks to all my family for their understanding and the love, support, and constant encouragement I have gotten over the years.

Finally, I would like to thank and dedicate this thesis to my father. It was you who always encouraged me for science. Although it has been two years since you have passed, I still take your lessons with me, every day.

Samin MOHAMMADI
04^{th} Dec 2018

Abstract

Nowadays, social media has widely affected every aspect of human life. The most significant change in people's behavior after emerging Online Social Networks (OSNs) is their communication method and its range. Having more connections on OSNs brings more attention and visibility to people, where it is called *popularity* on social media. Depending on the type of social network, popularity is measured by the number of followers, friends, retweets, likes, and all those other metrics that is used to calculate engagement. Studying the popularity behavior of users and published contents on social media and predicting its future status are the important research directions which benefit different applications such as recommender systems, content delivery networks, advertising campaign, election results prediction and so on. This thesis addresses the analysis of popularity behavior of OSN users and their published posts in order to first, identify the popularity trends of users and posts and second, predict their future popularity and engagement level for published posts by users.

To this end, i) the popularity evolution of ONS users is studied using a dataset of 8K Facebook professional users collected by an advanced crawler. The collected dataset includes around 38 million snapshots of users' popularity values and 64 million published posts over a period of 4 years. Clustering temporal sequences of users' popularity values led to identifying different and interesting popularity evolution patterns. The identified clusters are characterized by analyzing the users' business sector, called *category*, their activity level, and also the effect of external events.

Then ii) the thesis focuses on the prediction of user engagement on the posts published by users on OSNs. A novel prediction model is proposed which takes advantage of Point-wise Mutual Information (PMI) and predicts users' future reaction to newly published posts. Finally, iii) the proposed model is extended to get benefits of representation learning and predict users' future engagement on each other's posts. The proposed prediction approach extracts user embedding from their reaction history instead of using conventional feature extraction methods. The performance of the proposed model proves that it outperforms conventional learning methods available in the literature.

The models proposed in this thesis, not only improves the reaction prediction models to exploit representation learning features instead of hand-crafted features but also could help news agencies, advertising campaigns, content providers in CDNs, and recommender systems to take advantage of more accurate prediction results in order to improve their user services.

Keywords

Online Social Networks, Machine Learning, Prediction, Popularity, Representation Learning, Data Mining

Résumé

De nos jours, les médias sociaux ont largement affecté tous les aspects de la vie humaine. Le changement le plus significatif dans le comportement des gens après l'émergence des réseaux sociaux en ligne (OSNs) est leur méthode de communication et sa portée. Avoir plus de connexions sur les OSNs apporte plus d'attention et de visibilité aux gens, où cela s'appelle la popularité sur les médias sociaux. Selon le type de réseau social, la popularité se mesure par le nombre d'adeptes, d'amis, de retweets, de goûts et toutes les autres mesures qui servaient à calculer l'engagement.

L'étude du comportement de popularité des utilisateurs et des contenus publiés sur les médias sociaux et la prédiction de leur statut futur sont des axes de recherche importants qui bénéficient à différentes applications telles que les systèmes de recommandation, les réseaux de diffusion de contenu, les campagnes publicitaires, la prévision des résultats des élections, etc. Cette thèse porte sur l'analyse du comportement de popularité des utilisateurs d'OSN et de leurs messages publiés afin, d'une part, d'identifier les tendances de popularité des utilisateurs et des messages et, d'autre part, de prévoir leur popularité future et leur niveau d'engagement pour les messages publiés par les utilisateurs.

A cette fin, i) l'évolution de la popularité des utilisateurs de l'ONS est étudiée à l'aide d'un ensemble de données d'utilisateurs professionnels 8K Facebook collectées par un crawler avancé. L'ensemble de données collectées comprend environ 38 millions d'instantanés des valeurs de popularité des utilisateurs et 64 millions de messages publiés sur une période de 4 ans. Le regroupement des séquences temporelles des valeurs de popularité des utilisateurs a permis d'identifier des modèles d'évolution de popularité différents et intéressants. Les grappes identifiées sont caractérisées par l'analyse du secteur d'activité des utilisateurs, appelé catégorie, leur niveau d'activité, ainsi que l'effet des événements externes.

Ensuite ii) la thèse porte sur la prédiction de l'engagement des utilisateurs sur les messages publiés par les utilisateurs sur les OSNs. Un nouveau modèle de prédiction est proposé qui tire parti de l'information mutuelle par points (PMI) et prédit la réaction future des utilisateurs aux messages nouvellement publiés. Enfin, iii) le modèle proposé est élargi pour tirer profit de l'apprentissage de la représentation et prévoir l'engagement futur des utilisateurs sur leurs postes respectifs. L'approche de prédiction proposée extrait l'intégration de l'utilisateur de son historique de réaction au lieu d'utiliser les méthodes conventionnelles d'extraction de caractéristiques. La performance du modèle proposé prouve qu'il surpasse les méthodes d'apprentissage conventionnelles disponibles dans la littérature.

Les modèles proposés dans cette thèse, non seulement déplacent les modèles de prédiction de réaction vers le haut pour exploiter les fonctions d'apprentissage de la représentation au lieu de celles qui sont faites à la main, mais pourraient également aider les nouvelles agences, les campagnes publicitaires, les fournisseurs de contenu dans les CDN et les systèmes de recommandation à tirer parti de résultats de prédiction plus précis afin d'améliorer leurs services aux utilisateurs.

Mots-clés

Réseaux sociaux en ligne, apprentissage machine, prédiction, popularité, apprentissage de la représentation, exploration de données

Table of contents

List of Figures

Chapter 1

Introduction

Nowadays, every aspect of human life has been widely affected by social media. An enormous amount of data is uploaded to Online Social Networks (OSNs) every day which is analyzed and employed to improve the user services provided by those networks. OSNs are used by *professional users*, who get benefits of social media to promote their business, products, company, and etc, and also used by individual users, who take advantage of social media for the personal purposes such as sharing the details of their daily life by posting photos, writing statuses, sharing visited location, and so on. Both of those users try to attract as much as users they can in order to bring more visibility to their page and their published posts. By increasing the number of followers and friends, the visibility of the posts and consequently user engagement increase. More user engagement on the posts leads to more popular posts. *Popularity* on social media is usually measured by the number of followers and friends for users, and the number of likes, comments, shares, tweets, and etc for posts.

1.1 User Popularity

Popularity is very important for users on social networks especially for professional users who are following more serious goals of increasing visibility, turnover, sells, and so on. Reaching those goals through social media can be directly affected by the number of followers. While the number of followers for regular users is not as much crucial as for professional users. Many of professional users are willing to spend a considerable amount of money to increase the popularity value, even through unusual ways such as buying likes from *like farms* [2] [3]. The number of followers of a page has been found to be one of the most positive correlated features linking candidates' fan pages to the number of their votes in elections [4] [5]. OSNs offer a huge platform for professional users (i.e. companies, politicians, celebrities, etc.) to increasingly attract followers and promote their goals as much as possible [6].

Meanwhile, Facebook as the most popular OSN with more than one billion subscribers

defines a specific type of account for professional users, called *FanPages*[1]. This type of account has several features that distinguishes it from regular accounts. If a user likes a page, it will be added to the interest list of the user's profile. Professional users from various categories can create *FanPages* on Facebook to interact with their fans and customers. Apart from the general static attributes such as the page description and category selected by the page owner, the main dynamic attribute for each page is the number of fans (N_f) who have liked the page. This metric is publicly available for each *FanPage* and considered as the main metric that shows the popularity of a *FanPage* [7]. Even in major political events such as US presidential election, the popularity metric in different social media is the main metric to compare different candidate success in their campaign.

Understanding the popularity evolution and identifying the most influential factors on its variation will help professional users to take significant decisions about their attitude and behavior on different social media. In the present thesis, popularity evolution of professional users is elaborately investigated in order to model the variation patterns of their popularity trends. Additionally, the influential factors, including the activity volume and external events, on popularity evolution are also comprehensively studied. The main contributions of user popularity section are:

i. Proposing a methodology of monitoring the popularity evolution of professional users on Facebook in a noticeably micro level which not only is novel but also applicable to different types of OSNs.

ii. Identifying two main groups of users including fan-attractors who grew their N_f by different patterns, and fan-losers, users with a noticeable drop in their popularity trend.

iii. Finding several influential factors on popularity trend of users when the activity level of users or being celebrity are positively correlated to the trend of the number of fans.

1.2 Content Popularity

Users are the main actors on social media whom engagement and reactions to the published posts play a substantial role in information propagation and popularity of the post [8] [9]. The total number of engagements on a post shows the number of reactors (who reacted to the post), also known as the popularity number. Predicting this value and its involved reactors are two significant prediction tasks, which supply valuable information for many applications such as providing better solutions for content placement in networks, more efficient advertisement campaigns, and providing accurate recommendations.

The first task, predicting popularity size, has been inspected by many researchers using structural, profile, network, and content features [10] and recently some temporal information about early reactors [11] [12]. However, it is not the focus of the present thesis. Our focus in this study is to predict future reactors who are going to react to recently published posts, which is the second mentioned task. There is a big number of studies that have

[1] http://www.facebook.com/about/pages/

investigated the prediction of future reactors using a combination of social and content features as well as temporal features [13] [14]. These models have to define features and extract the most suitable and efficient ones manually. Where feature extraction is a very difficult and frustrating task. Instead, we avoid this task in our study by taking advantage of feature embeddings.

Feature embeddings are basically anything that can act as a hidden representation for a given object. Embedding means converting data to a feature representation where certain properties can be represented by notions of distance. It is essentially projecting the data to a high dimensional feature space, so that the features that are more or less alike have a small distance between them in the embedded space. We propose a model to derive user embeddings and use them to predict future reactors. Therefore, the model will not need feature selection step any more. The prediction is done using two proposed prediction methodology. First, a method based on Point-wise Mutual Information (PMI) inspired by the $Word2vec$ language model [1]. And second, a model based on user embedding features derived from the Word2vec model. Both of the models take users' reaction history as input. The main contributions of reactors prediction section are:

i. The proposed models do not need to manually select and manage features to predict future reactors, which is the novelty of the study.

ii. Our model requires minimum data for prediction as input, which is users' reaction history, in compare to similar models that required hand-crafted features.

iii. The models proposed in this study are general and can be applied to any social networks data.

Thesis Structure

The remainder of the thesis is organized as follows: The second chapter presents the background related to the user and content popularity prediction models on social media and discusses the latent representation learning methods. In the first section of the third chapter, we will present our study of popularity evolution on Facebook professional users. First, we discuss popularity analysis from different perspectives and then present the classification model and evaluation results. Finally, we conclude our study and discuss the potential future works. The second section of Chapter three extends the first section's study by expanding the behavior analysis to a long-term evolution of user popularity. Indeed, this section focuses on the popularity variation of the same users in a longer period and presents the influential factors and the impact of external events on the user's popularity pattern.

In Chapter four, we present our study on future likers prediction. We first discuss our methodology and prediction model. Then, we explain the evaluation of the proposed method and analyze the results. Chapter five presents the reactions prediction model using embedding features. After presenting the proposed method, it explains the dataset used to evaluate the model. Then, we discuss the results and the performance of the model in compare to some baseline methods. The final section concludes the thesis by summarizing our main proposals and contributions and presents some perspectives of our work.

Chapter 2

Related Works

Contents

2.1 Introduction

Popularity is one of the most well-studied aspects on social media [15] [16] where it has become one of the main utilities that is used in advertisements, marketing, and predictions [2]. The term *'popularity'* refers to different metrics such as the number of likes, views, or votes that a page or a content receives [15] [17]. Studying the popularity pattern and prediction of its future behavior are interesting research topics which can be investigated about both users [18] and contents [19].

2.2 User Popularity

Popularity of users on social media is measured by the number of their followers and friends, as well as the number of user engagement on their published posts. The variation of populairty and consequently its behavior is used as a criterion to make a policy in order to succeed in attracting as many followers as possible. Barclay *et al.* [4] investigated the correlation between political opinions on Facebook and Twitter in the US presidential elections of 2012. They showed that the number of fans and the sentiment of comments are the most-correlated features to the candidates final votes. In another similar work, Barclay *et al.* [20] demonstrated the number of likes of the Facebook *FanPage*s of the parties as a predictor of election outcomes with 86.6% accuracy.

Meanwhile, a number of studies have focused on identifying the influential factors on attracting new fans and increasing user engagement level [21] [22]. Authors in [23] performed an empirical study on a sample of posts created by different brands on their Facebook *FanPages*. The work consists in conducting a content analysis in order to relate the characteristics of the posts to users' engagement. They investigated the impact of some factors such as emotion and testimonial presence. Pronschinske *et al.* [24] studied the relationship between the attributes of Facebook pages and the number of page likes, where the number of page likes is the page's popularity. They showed that being authentic by indicating a page as an official page and linking a website to a Facebook page as well as having more engagement in the posts of a page will attract more fans.

With assumption of the positive influence of popular reviewers on the final popularity of a product, authors in [25] made a model using machine learning techniques to classify reviewers into high/low popularity based on their profile characteristics. Based on this work, businesses can identify potentially influential reviewers to request them for reviews in order to increase the popularity of product. Ferrara *et al.* [26] measured the total number of likes and comments received by a user's media in order to investigate the popularity of users. The study also accounts for the total number of times a user likes or comments someone else media, namely the number of social actions that this user performs. Comparing social actions users popularity distributions shows that the social actions distribution is broad but with a steeper slope than users popularity. This implies that there exist relatively less users (with respect to the popularity distribution) who produce many likes or comments to others' media.

Analyzing the influence patterns among Twitter users, and understanding how the users

considered as experts in a given field in order to promote the growth of the number of followers of other users is the aim of authors in [27]. Their result proves that it is not easy to determine the factors that allow a Twitter user to get more followers. They also observed that users tend to keep steady, and it is not very frequent that a user changes linguistic category. It is worth mentioning that several companies monitor Facebook *FanPages* activities and provide reports, by charging their customers, with general analysis for their clients. One of them that provides aggregated popularity results for single users, is *SocialBakers*. They claim that their services allow brands to measure, compare, and contrast the success of their social media campaigns with competitive intelligence.

In summary, although few studies have looked to the different aspects of Facebook *FanPages*, but their focus were mostly for a small group of users. To the best of the authors knowledge, our study is the first one that has specifically investigated the evolution of popularity in a large scale and for a long period. The study looks to this aspect in detail for a list of 8K popular *FanPages* and also investigates the influential factors to the popularity evolution trends.

2.3 Content Popularity

From the content perspective, once a content is published on a social network, it attracts different amount of users interactions depending on its interestingness, topic, publisher's reputation, published time and etc. [28] [29]. Meanwhile, some contents succeed to attract more user engagements and become popular [30]. Popularity of a content usually assesses by different cascading metrics such as number of likes, shares, views, etc.

Simultaneously, many studies have tried to model and forecast popularity of content on social media [15]. Bandari *et al.* utilized article features like source, category, and subjectivity to predict the popularity of an article on Twitter with 84% accuracy. Lerman *et al.* used a stochastic model to predict how popular a newly posted story will be based on the early reactions of Digg users [31]. In [32] and [33] researchers used temporal content features to predict the popularity of content by exploiting time series clustering techniques and linear regression methods. Different categories of features have been examined to predict the popularity of content [10] and in [11] temporal features are illustrated as the best predictors.

Cvijikj *et al.* [17] analyzed the effects of content characteristics on user engagement on Facebook *FanPages*. They found that providing informative and entertaining content significantly increases the user's engagement level. To enhance the number of likes and comments of a post, Vries *et al.* [16] found that highly vivid and interactive posts like videos and questions can attract more likes and comments than other kinds of post. To predict a concise popularity score of social images, visual sentiment features are used together with context features [34]. Experiments on large scale datasets show the benefits of proposed features on the performance of image popularity prediction. Moreover, their qualitative analysis shows that sentiments seem to be related to good or poor popularity. In another research [35], authors study the effects of visual, textual, and social factors on popularity in a large real-world network focused on fashion. They found significant statistical evidence

that social factors dominate the in-network scenario, but that combinations of content and social factors can be helpful for predicting popularity outside of the network. Their in depth study of image popularity suggests that social factors should be carefully considered for research involving social network photos.

Swani *et al.* [36] investigate the key factors that contribute to Facebook brand content popularity metrics (i.e., number of likes and comments) for Fortune 500 companies' brand posts in business-to-business (B2B) versus business-to-consumer (B2C) markets. The results indicate that the inclusion of corporate brand names, functional and emotional appeals, and information search cues increases the popularity of B2B messages compared with B2C messages. Moreover, viewers of B2B content demonstrate a higher message liking rate but a lower message commenting rate than viewers of B2C content. Predicting the trend of popularity for a content (which can be a text, video, or image) and more importantly identifying the users who are going to react to that content are very valuable information for different entities such as service providers to rank the content better [37], to early discover trending posts, to improve recommendations and even to improve their content delivery networks and user experiences [38]. This kind of prediction tasks are mainly based on the features of contents and early adapters. Depending on the social network's type, adapters can be interpreted as either likers, resharers, viewers, or so on. In [39], popularity of a content is predicted using the structural diversity of early adapters. In other studies, temporal features of early adopters are realized as the most predictive features among different features of content, user and network [10] [11] [40].

Looking at the models that have been developed on different content popularity prediction tasks on OSNs shows that most of them focused on predicting the popularity size of contents in future. There are very rare researches on identifying the users who are going to react to the contents published on OSN in future [41]. Although, interactors prediction on OSNs is somehow similar to well-studied rate prediction on recommender systems (RS), but there is a main differece which makes RS models improper to apply directly on interactors prediction on OSNs. Rate prediction models on RS are mainly based on interest, whereas OSNs models are primarily based the mixture of friendship and interest. Petrovic *et al.* [42] tried to predict interactors using a machine learning method based on the passive-aggressive algorithm.

Authors in [13] have proposed a tree-structured Long Short-Term Memory (LSTM) network to learn and predict the entire diffusion path of an image in a social network. By combining user social features and image features, and encoding the diffusion path taken thus far with an explicit memory cell, the model predicts the diffusion path of an image. In [14], authors investigate the sequential prediction of popularity by proposing a prediction framework, by incorporating temporal context and temporal attention into account. Our study is different from the mentioned and other similar studies [43] [44] because of focusing on only users latent likelihood extracted from their reaction history.

2.4 Latent Representation on OSNs

In previous studies, different learning methods have been used to model social networks prediction tasks [19]. Recently, deep learning methods attract attention of researchers in variety of studies including the prediction tasks through OSNs. These methods have achieved more vivid and appreciated results in comparison to the conventional learning methods [45] [46] [47]. One of the successful deep learning architectures is word2vec model to capture word embeddings in natural language processing applications [1]. It extracts words semantic similarity using a simple architecture.

2.4.1 Pointwise Mutual Information on OSNs

Point-wise mutual information (PMI) is a measure to model the dependency of two instances of random variables used widely in information theory, Natural Language Processing (NLP), Recommender Systems (RS) and OSNs. NLP models use PMI to find the strength of association between words [48] [49] [50]. In [51], PMI is used to compute semantic similarity and relatedness of words where it achieves outperforming results. RS also take advantage of PMI as one of the measures which used to find users and items similarities [52] [53]. Kaminskas *et al.* used PMI between items to measure surprise in RS and compared its results with a content-based surprise measurement [54]. In [55], authors get profit of PMI between different recipes' ingredients and predict recipe ratings given by web users. Spertus *et al.* compared different similarity metrics including PMI to compute similarity of Orkut communities in order to find users' interesting communities and exploit them in a recommendation task [56].

Social networks applications also benefit from PMI and use it for two primarily objectives, first word and consequently content similarity, and second user similarity. Different problems have been studied on OSNs using words' PMI metric such as content sentiment analysis, topic detection, content classification and so on [57] [58], but our focus in this study will be on users similarity. Authors in [59] exploited PMI to measure the network similarity of users based on their mutual friends on social networks. Following the aim of this study, we use PMI between users to find their interaction similarities. Our proposed model is inspired by Word2vec language model [45] to compute users co-occurrences. Akin to Wodr2vec which extracts word-context pairs from sentences considering a *window* of size w, our model will also employ the idea of *window* and consider each user to be paired with w users before and after that user in the like streams. More detail will be provided in Section 4.3.1.

2.4.2 Word2vec

Authors in [1] have proposed two architectures as two new neural word embeddings structures. The first is Continuous Bag-of-Words (CBOW), which predicts the current word based on the context, and the second approach is Skip-gram which predicts surrounding words given the current word. Skip-gram with negative sampling (SGNS), also known as Word2vec (a sample is shown in Figure 2.1) is an efficient method for learning high-quality

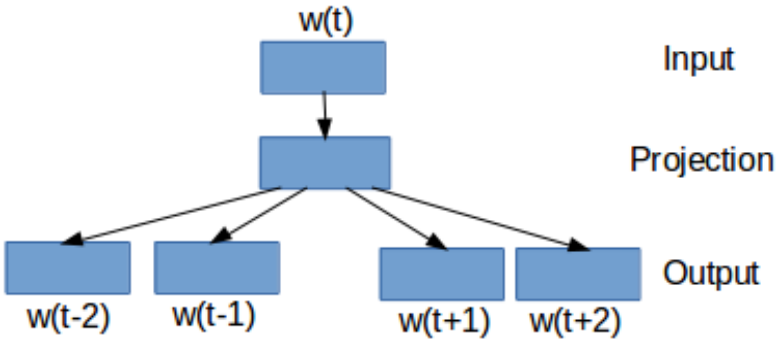


Figure 2.1 – The Skip-gram model architecture for predicting surrounding words given the current word *w(t)* [1].

word representation that captures the semantic relation of a word with its surrounding words in a corpus [45].

The Skip-gram approach trains high quality word vectors using a simple architecture. As shown in Figure 2.1, the model predicts the surrounding words $(w(t-2), w(t-1), w(t+1), w(t+2))$ given the current word $w(t)$. The goal is to find word vector representations that help to predict the nearby words. More formally, given a sequence of words $w_1, w_2, ..., w_k$, where $w_i \in W(the\ vocabulary)$, the goal is to maximize:

$$1/k \sum_{i=1}^{k} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{i+j} \mid w_i)$$

Noticeably improved results from using Word2vec can be seen not only in Natural Language Processing (NLP) [1], but also in other research domains such as social networks [60] [61]. In a method inspired by Word2vec [61], authors exploit the same framework of neural word embedding and produce embeddings for items in an item-based collaborative filtering. DeepWalk [60] uses random walks like the sequences of words in Word2vec to learn the latent representations of users to use on multi-label network classification tasks. Node2vec [47] is a successful method to extract user embeddings where it maximizes the likelihood of preserving network neighborhoods of nodes. As words in sentences and users in random walks correspond to users in interaction sequences, it allows us to utilize a similar method to derive users' reactions likelihood from their interaction sequences. We explain next how this adaptation has been done in this study.

Part I
User Popularity on Social Media

Chapter 3

Popularity Evolution of Professional Users

Contents

Subpart A

Identifying User's Popularity Behavior on Facebook

3.1 Abstract

Popularity in social media is an important objective for professional users (e.g. companies, celebrities, and public figures, etc). A simple yet prominent metric utilized to measure the popularity of a user is the number of fans or followers she succees to attract to her page. Popularity is influenced by several factors which identifying them is an interesting research topic. This study aims to understand this phenomenon in social media by exploring the popularity evolution for professional users in Facebook. To this end, we implemented a crawler and monitor the popularity evolution trend of 8k the most popular professional users on Facebook over a period of 14 months. The collected dataset includes around 20 million popularity values and 43 million posts. We characterized different popularity evolution patterns by clustering the users' temporal number of fans and study them from various perspectives including their categories and level of activities. Our observations show that being active and celebrity correlate positively with the popularity trend.

3.2 Introduction

In the fast-paced digital world, Online Social Networks (OSNs) have experienced a massive growth in their variety and usage over the past decade. These systems offer a huge opportunity for professional users (i.e. companies, politicians, celebrities, etc.) who aim to both attract new followers and interact better with them [6]. Facebook as the most popular OSN with more than one billion subscribers defines a specific type of account for professional users, called *FanPages*[1]. This type of account has several features that distinguish it from regular accounts. If a user likes a page, it will be added to the interest list of the user's profile. Professional users from various categories can create *FanPages* on Facebook as a means of interacting with their fans and customers. Apart from the general static attributes such as the page description and category selected by the page owner, the main dynamic attribute for each page is the number of fans (N_f) who have liked the page. This metric is publicly available for each *FanPage* and considered as the main metric that shows the popularity of a *FanPage* [7]. Even in major political events such as US presidential election, the popularity metric in different social media is the main metric to compare different candidate success in their campaign.

Several studies have emphasized the role of N_f as a comparative and competitive metric for professional users. Many of professional users are willing to spend a considerable amount of money to increase this value, even through unusual ways such as buying likes from *like farms* [2] [3]. The number of likes of a page has been found to be one of the most positive correlated features linking candidates' fan pages to the number of their votes in elections [4] [5]. Attracting Facebook fans is also used as a marketing strategy [62] and provides a metric to measure the return on social media investment [63]. We will use the term *popularity* to refer to the number of likes of a page. To the best of authors' knowledge, even though a number of papers have studied the popularity trends of content

[1] http://www.facebook.com/about/pages/

and posts [16] [64], there is no study on evaluating the popularity evolution of users, especially by the focus on professional users.

This section of research studies the temporal popularity evolution of professional users through their *FanPages* on Facebook and attempts to identify the factors that influence the popularity trends. The objectives pursued here are designed to answer the following research questions:

i. How does the temporal popularity of users vary overall and in accordance with users' business sector (Facebook pre-defined categories)?

ii. What temporal patterns can be identified from the time-series N_f of pages?

(*iii*) What are the factors influencing the popularity trends?

To answer the stated questions, an extensive list of the most popular professional users in terms of N_f was selected and the required data collected by implementing advanced data collection tools. Our dataset includes 8K of *FanPages* that have the highest number of fans validated by a third-party portal *Social Bakers*[2].

The main contributions of this section are:

i. The proposed methodology of monitoring the popularity evolution of professional users on Facebook in very micro level is novel which is applicable to different types of OSNs.

ii. Following the methodology, we classified the users in two main groups: First, fan-attractors who grew their N_f by different patterns, and second, fan-losers, users with a noticeable drop in their popularity trend.

iii. We found several influential factors on the popularity trend of users. The activity level of users or being celebrity are positively correlated to the trend of the number of fans.

The rest of this section is organized as follows: We present related work in Section 2 followed by Section 3.9 describing the methodology and the dataset. Section 3.10 represents a general overview of the popularity and its evolution. The model and results are discussed in Section 3.11 and finally Section 3.6 concludes this study.

3.3 Data Collection and Dataset

The objective of this study is to explore how the popularity of top professional Facebook *FanPages* evolves. To this end, we first selected 8K of the top Facebook *FanPages* based on their N_f from the previously mentioned third-party application *Social Bakers* which ranks users based on the number of fans.

In order to monitor the popularity evolution of the selected users and generate a time-series of their N_f and of their activities; we implemented three crawlers as follows: Firstly,

[2]http://www.socialbakers.com/

Table 3.1 – Dataset Characteristic

Attribute	Value
Duration	14 months
Crawling Period	Sep'13 - Oct'14
#Sample per day	6 snapshots (Q4h)
#Users (#*FanPages*)	7,875
Total #Samples in dataset	20M samples
Avg(#Sample) per user	1,298 samples
Median(#Sample) per user	1,297 samples
Total #Post in dataset	43M posts
Avg(#User_Post) per month	107 posts
Median(#User_Post) per month	24 posts

we implemented a data collection tool that queries FB public API to collect the number of fans. The data collection is performed for the selected 8K users over a period of 14 months from September 2013 to October 2014. To have enough detail, the value of N_f is recorded, every 4 hours (6 times per day). The second crawler collects the general information of users from their profile which includes detailed information such as their pre-defined categories, description of the page, and etc. The third crawler collects the activity (published posts) of users and its associated attributes on the period of our study. A summary of dataset's main characteristics is presented in Table 3.1.

3.4 Evolution of Popularity

Before clustering, we go through the analyzing aggregated popularity evolution of users to provide an insightful vision of the dataset. During the initial analysis, a group of users is identified who have a sudden and large peak in their N_f in a very short period of time. By looking carefully to their data, we found that this peak reflects the impact of a newly announced service by Facebook, named *GlobalPage* [65]. Facebook *GlobalPage* is a new page structure for big brands which are active across globe and have several separate pages with the same name but active in different languages and different locations. These pages which formed almost 10% of the dataset, were excluded from it because their trend are not aligned with the aim of this study which is to identify real popularity trends and their effective factors.

3.4.1 Popularity Analysis - In Overall

Monthly popularity value is defined indicating the average value of user's N_f in each month. Since our dataset covers 14 months, each user has a 14-entries vector representing her popularity trend in the period of the dataset.

By considering the overall changes in N_f from M1 to M14 for each user, despite the probable peaks and drops, 80% (5798 out of 7216) of the users attracted new fans and on the other hand 20% (1418 out of 7216) lost fans during this 14-month period. Figure 3.1 shows the distribution of users' popularity from the first month (M1) to the last (M14). The median values for M1 and M14 distributions are 1.3 and 1.7 Millions respectively,

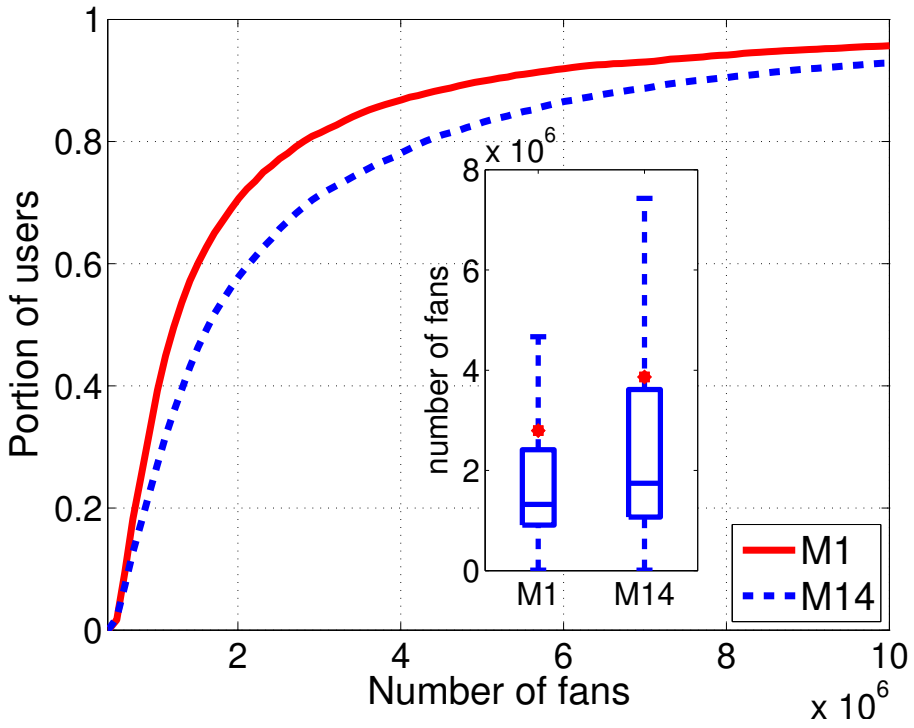


Figure 3.1 – CDF (and boxplot with red dot representing the Mean value) of the N_f of users in M1 and M14

which this median value increased from M1 to M14 by 30% (and 38% increment for mean value).

Figure 3.2 represents the distribution of users based on the percentage of their N_f growth during the period of this study. As shown in the figure, the growth rate of the number of fans for pages who lost fans is not less than -20% and the major range of fans lost are between -5% and 0%. On the other hand, most of the fan-attractor pages are in the range of 10% to 30% growth and the distribution continues in a long-tailed pattern.

3.4.2 Popularity Analysis - Category Wise

Each page is assigned to a business sector by the page owner in the time of subscribing called category. To investigate the users' distribution and overall popularity evolution inside the categories, we chose 17 (out of 158) categories those that include more than 1% of the total pages in the dataset separately and more than 75% in sum shown in Table 3.2 The main observations from Table 3.2 are as follow:

i. *Musician Band* is the most populated category in our dataset which shows users in this category are the most popular ones in the dataset.

ii. The percentage of average growth in the fifth column refers to the average N_f growth of users in each category over 14 months. Interestingly, it shows that the *Athlete,*

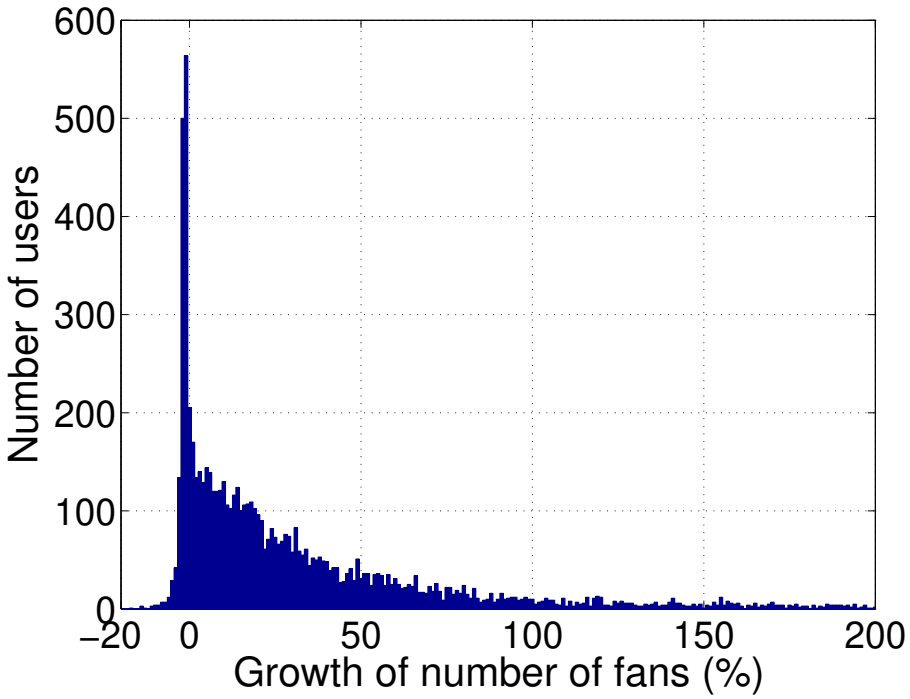


Figure 3.2 – Distribution of users based on the percentage of their N_f growth, during 14 months (from M1 to M14)

Actor Director, and *Sports Team* categories have the highest percentage of growth, and on the contrary *Community* has the lowest. This indicates that users in the three mentioned categories are successful in attracting new fans on average, whereas *Community* category users show a negative growth.

iii. The last column of the table shows the users' median value of the N_f growth in each category. A negative value here shows users of that category are loosing fans which means people unfollow the pages by *unliking*. *Community* is the only category which has negative median growth. This means that most of the users in this category have lost some of their fans.

3.5 Users' Clustering

This section aims to analyse the popularity in the user level and try to identify different clusters of users with similar patterns in their popularity trends. To this end, the evolution of N_f is modeled by exploiting different clustering techniques and investigating different characteristics (popularity range, category and activity distributions) in each identified cluster.

Table 3.2 – Populated categories distribution in the dataset. Fifth and sixth columns indicate the growth rate of average and the median of N_f over 14 months respectively.

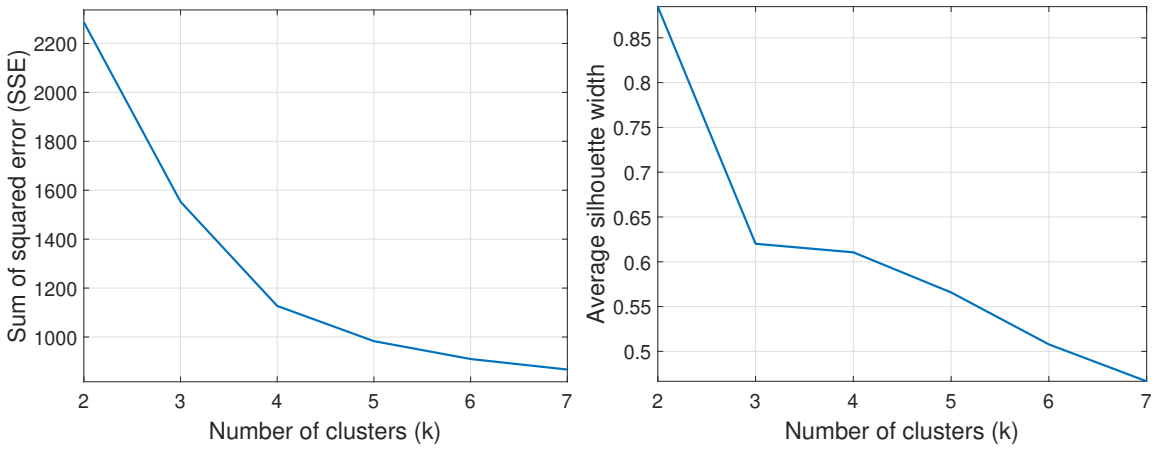
#	FB Category	#Pages	%Pages	%Avg. growth	%Median growth
1	Musician Band	1231	**17**	47	32
2	Community	986	13.7	**2.1**	**-1.5**
3	Tv Show	477	6.6	53	15
4	Movie	413	5.7	28	18
5	Food Beverages	302	4.2	19	11
6	Product Service	267	3.7	24	15
7	Public figure	246	3.4	64	33
8	Company	188	2.6	23	15
9	Athlete	188	2.6	**101**	65
10	Actor Director	179	2.5	**97**	50
11	Entertainment	166	2.3	26	4
12	App page	143	2.0	17	8
13	Clothing	139	1.9	29	19
14	Media News	134	1.8	76	42
15	Sports Team	125	1.7	**92**	60
16	Games Toys	109	1.5	13	6
17	Health Beauty	85	1.2	17	7

3.5.1 Feature Vector and Clusters

To cluster users based on the popularity attributes, a 14-entry monthly popularity vector for each user is used as a feature vector in the clustering method. The entries represent the monthly N_f of users that have values over the range of one hundred thousand to one hundred million. The goal is to group the users with similar popularity evolution into a cluster, regardless of the value of N_f. To clarify this point, consider two *FanPages* from quite different ranges of popularity, which both have 50% growth of N_f with the same trend over the same time period. They should be assigned to a same cluster because their popularity trend are similar. To this end, we used the Min-Max normalization method which scales every feature vector into $[0, 1]$ by obtaining the values 0 and 1 at the minimum and maximum points, respectively. The feature vectors thus represent the time-series popularity trends of users.

Next we applied several clustering algorithms including K-means [66], KSC [67] and K-shape [68] and as the outcome of all of them were similar, we consider the K-means clustering algorithm to the above-mentioned feature vectors. K-means requires the number of clusters (k) as the input parameter. There are different approaches to detect the optimal number of clusters. In this study, we used the elbow method [69], which considers the within-cluster sum of the squared errors (SSE) to find the proper k for our dataset. Figure 3.3a shows the SSE results for different k numbers applied to the dataset

As depicted in Figure 3.3a, the distortion of SSE goes down rapidly by increment of



(a) SSE for different cluster numbers (b) Silhouette width for different cluster numbers

Figure 3.3 – SSE and Silhouette width test to find the proper k value for the dataset

k to the value of 4. Then it descends slowly to 5 and continues with slower decrement. It seems that the diagram reaches an elbow at $k = 4$. However to be more assured of an appropriate k value, the Silhouette width [70] of different k values is also computed. The concept of silhouette width involves the difference between the within-cluster tightness and the separation from the rest of clusters.

Figure 3.3b shows the average Silhouette width for different numbers of cluster. The average Silhouette width is almost constant with k increasing from 3 to 4. This means that with k equals to 4, users are located in as right cluster as with 3. But as the SSE in Figure 3.3a has an impressive decrease with 3 clusters, we chose 4 as the proper number of clusters.

Figure 3.4 represents the normalized popularity trends for the clusters. Each plot shows the average value of the normalized N_f belonging to the users in one of the cluster. In general, three of the identified popularity patterns are ascending by means of different behaviors, and one of them is descending. In summary we can observe the following points:

i. Users are continuously losing their fans in the first cluster (*Cluster-1*) which includes 20% of our dataset population.

ii. The most populated cluster is the *Cluster-2* by 43% of the users. It shows an ascending popularity growth behavior in average. This means that the popularity of the users in this cluster is constantly increasing due to attracting new fans.

iii. *Cluster-3* has 13% of the dataset population and users in this cluster show a sudden growth (around 80%) in the first half of the time and then their growth is stopped and somehow saturated in the second half.

iv. *Cluster-4*, with 25% of the users, shows an opposite behavior to *Cluster-3*. Its users show near to 30% growth in the first 7 months and then 70% during the last 7 months.

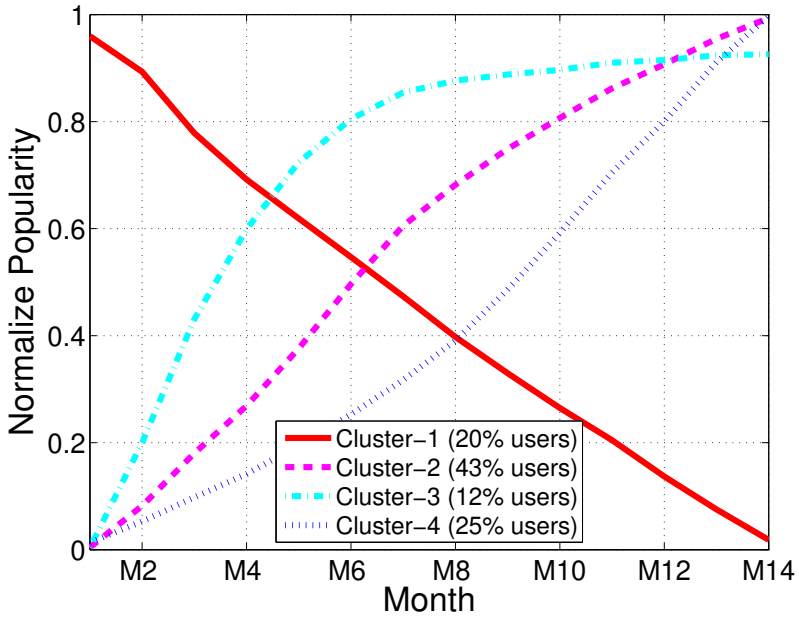


Figure 3.4 – Normalized popularity trends of four clusters

Next we characterize the identified clusters from three perspectives, their popularity, category and activity.

3.5.2 Popularity Distribution in each Cluster

This section analyzes the clustering results with respect to the users' popularity distribution. The aim is to identify how the normalized popularity trend can be affected by the absolute value of N_f. First we look to the distribution of popularity in the clusters. Figure 3.5 shows the CDF plots of the last month (M14) users' popularity in four identified clusters. The first interesting point in this figure is the popularity distribution of users in *Cluster-1*. As we saw earlier in Figure 3.4, users in this cluster are gradually losing their fans. Figure 3.5 shows most of these users are less popular than the users in other clusters. Almost 65% of them have less than 1M fans, and the number of users which have more than 2M fans does not exceed 10%.

According to this plot, three other clusters include users with much higher values of N_f. It can be observed that users in two of the most fan-attractor clusters (*Cluster-2* and *Cluster-4*) are more popular and have high N_f in compare to users in the other two clusters. The median values of popularity in these two clusters are almost 2M fans. While only 30% and 10% of users in *Clusters 3* and *1* have more than 2M fans.

Thus, the most popular users belong to *Cluster-2* and *Cluster-4*, which both represent exclusively fan-attractor behaviour. In contrary, most of the less popular users are in *Cluster-1* and *Cluster-3*, where their popularity pattern show a fan losing behavior or of being almost saturated. To conclude this section, in general more popular users show very

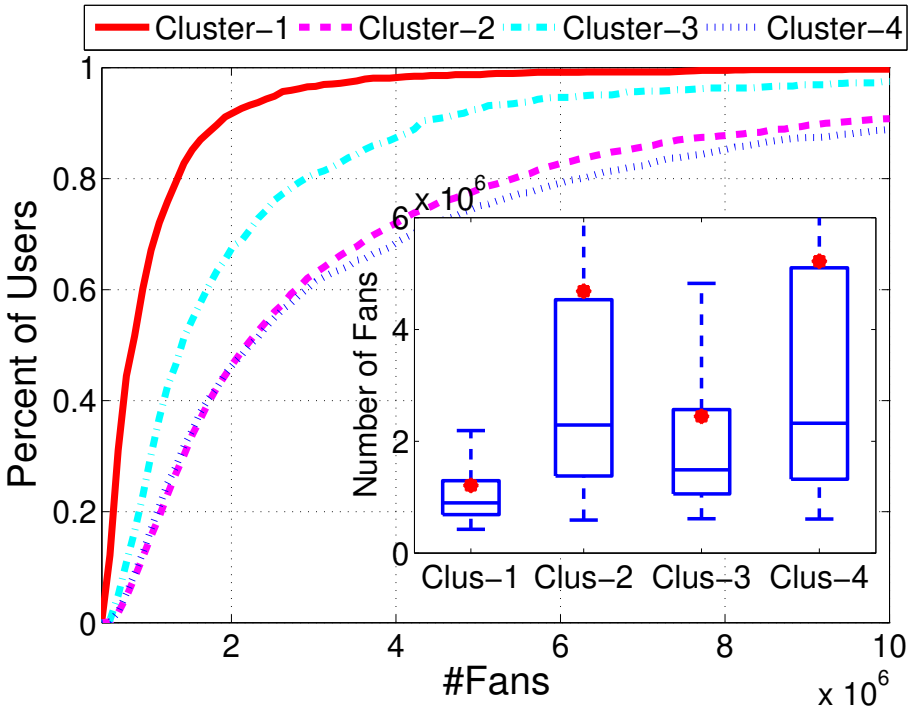


Figure 3.5 – CDF (and boxplot) of the distribution of users N_f in four identified clusters. (the red dot inside boxplot represents the Mean value of the distribution).

sharp fan attracting trends while less popular ones show fan losing or saturating trends.

3.5.3 Category Analysis

In this part we investigate the distribution of categories inside the identified clusters to understand if there are categories with a dominant population in a specific cluster. Figure 3.6 shows the distribution of the 17 most populated categories, mentioned earlier in Table 3.2, across the identified clusters.

An interesting observation from the category distribution is the high presence of the *Community* and *Entertainment* categories in *Cluster-1*, with around 85% and 40% portion of presence, respectively. Given that the users in this cluster are losing their fans, and the *Community* category is the second most populated category with 13.7% of the users in the dataset, it can be concluded that it is also the biggest set of fan-loser users. According to the Facebook[3], "a Community Page is a page about an organization, celebrity or topic that it does not officially represent. It links to the official page about that topic." Our observations show that a *Community* page is a place that Facebook users gather to share their ideas, images, posts around a specific topic, company, or celebrity and cannot remain attractive to users over time. One of the reason we found is the new feature of Facebook "Verified" which provide the possibility for verifying popular pages which Facebook started

[3]https://www.facebook.com/help/187301611320854/

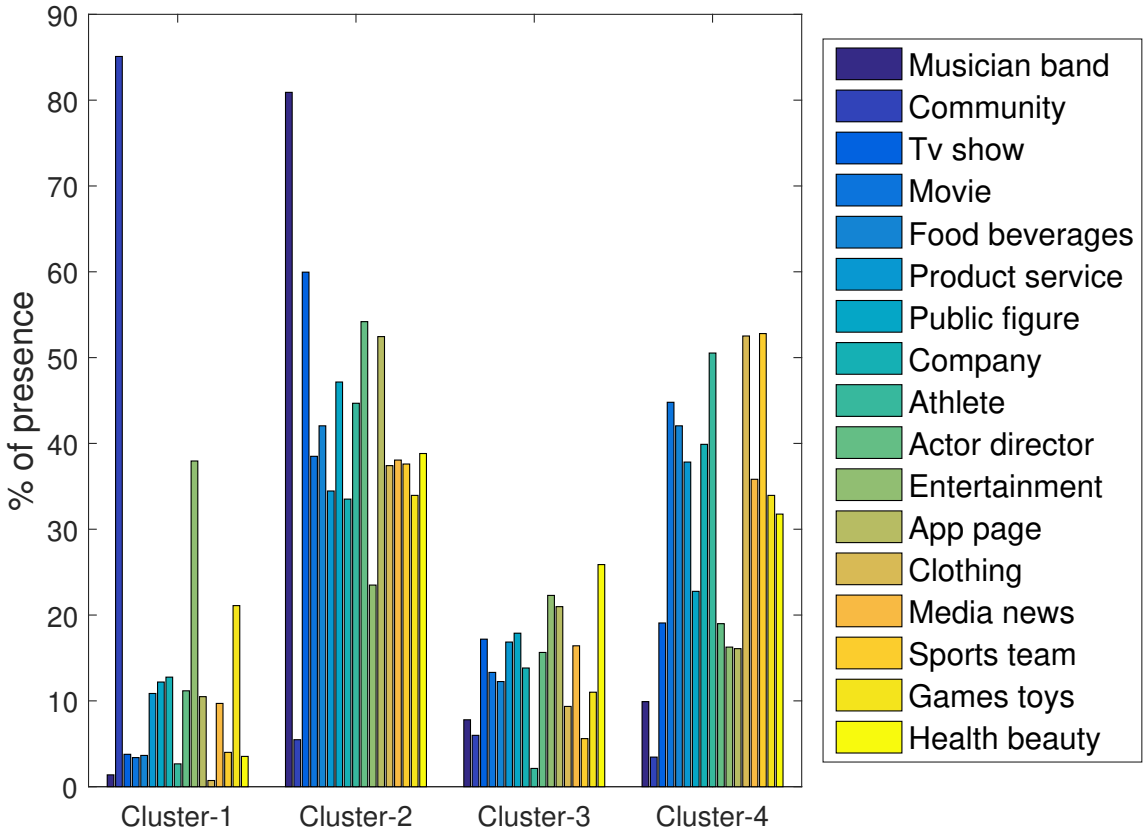


Figure 3.6 – Distribution of predefined Facebook categories in each identified cluster

in May 2013. After verification, people are more likely following the verified pages instead of the community pages.

In summary, according to the popularity trend of other three clusters and category distributions of *Cluster-1*, we can say more than 80% of users from all categories except *Community* and *Entertainment* categories are attracting new fans.

Cluster-2, which shows a fixed rate of popularity growth, includes a high presence of *Musician band* and *TV show* categories, which are two of the three most-populated categories with 17% and 6.6% of the users in the dataset. These two categories, accompanied by *Actor director*, contain most of the celebrities' pages in our dataset. On the other hand, as *Cluster-2* shows the most successful fan-attracting trend, we can indicate that the pages of celebrities are always interesting for people to follow. Around 30% to 50% of other categories' users also show similar pattern of attracting new fans. The distribution of categories in *Cluster-3* shows almost an equal presence of all categories without any dominant one, except a minimum presence of *Athlete* categories. The trend of this cluster could have different explanations like fan-saturation, reduction of the activity or external events which have the same side effect on users in different categories. In the next section, we look for the effect of activity volume on users' fan-trends as a probable influential factor.

Cluster-4, which includes 25% of our users, has a variety of categories distribution. Three categories, *Athlete*, *Clothing*, and *Sport team* have more than 50% of their population in this cluster. According to the popularity pattern of this cluster, most of the users experienced more than 70% of their popularity growth in the second half of the study period. Some famous celebrities such as *Neymar* (Football player), *Real Madrid C.F.* (Sport team) are in this cluster. For users such as those related to football, the most probable reason of significant N_f growth may be the main events of European leagues which are overlapped with the second half of our dataset period.

As a summary of this part, we saw that *Community* is characterized as the most fan-losing category with a major presence in *Cluster-1*. The categories containing more celebrities are the most fan-attracting ones, with a significant presence in the two most fan-attractor clusters, *Cluster-2* and *Cluster-4*.

3.5.4 Activity Analysis

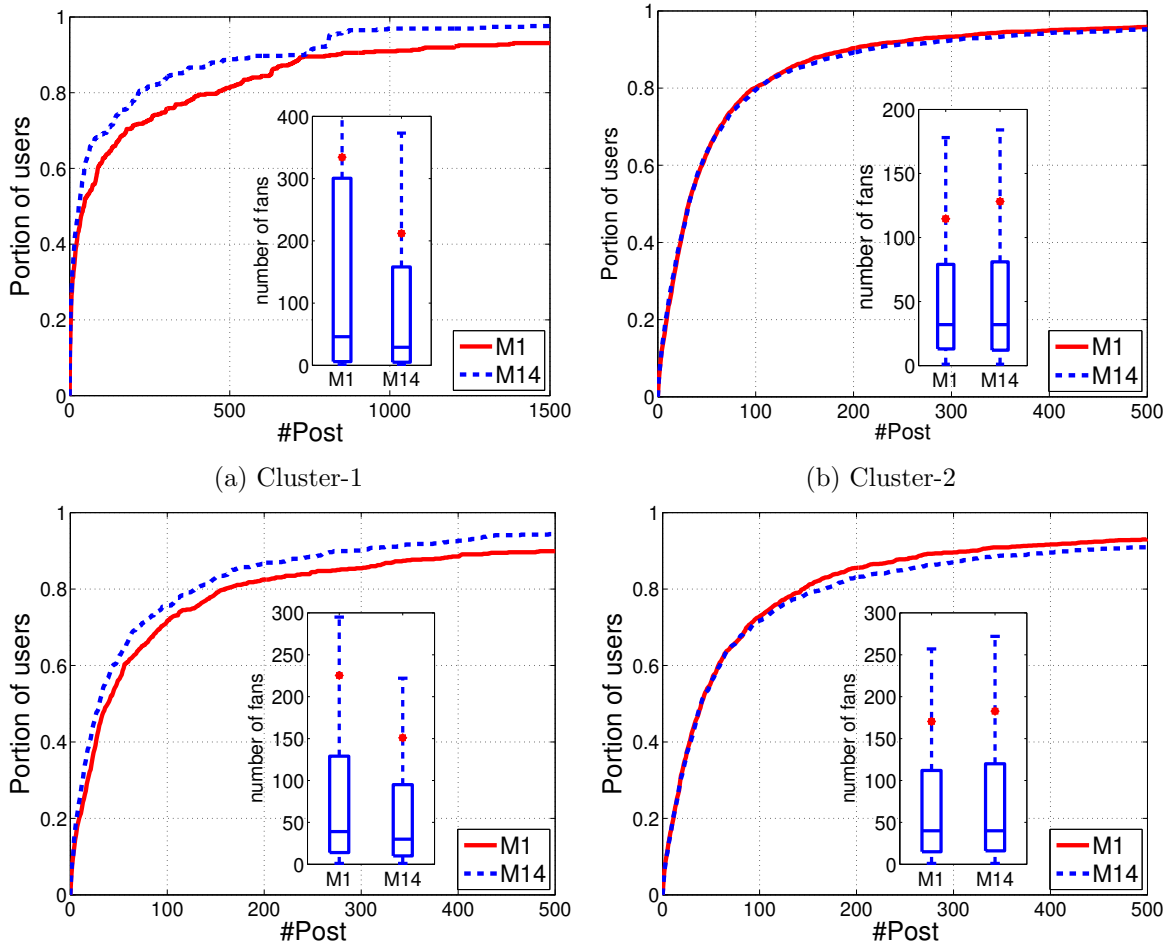
Being active in Facebook by continuously publishing new posts, can ensure professional users to stay in touch with their followers and attract new ones as well [21]. To understand the impact of activity on popularity, Figure 3.7 shows the CDF plots of the number of published posts by users in four clusters for M1 and M14. It illustrates that the published posts of the users in *Cluster-1*, who lost their fans, declined from M1 to M14. This can be observed for the distribution of users in *Cluster-3* as well (Figure 3.7c). As discussed before, the N_f of users in this cluster is almost constant for the second half of the study period. It can be concluded that the reduced number of activity in these two clusters is an important factor for the lost of fans in *Cluster-1* and the failure to attract new ones in *Cluster-3*.

In contrast, the activity level of users have not changed substantially in the two most fan-attracting clusters, *Cluster-2* and *Cluster-4*. Even we can see a small increment in the activity curve of *Cluster-4*; the number of users who published more than 150 posts in the last month is greater than the number of users who posted that much in the first month. Considering their popularity trends which show a continuous growth, it can be deducted that being constantly active effect the process of attracting new fans.

In a nutshell, we observe that staying active in terms of publishing posts can help to attract new fans and followers whereas reducing the activity level can lead to stagnant number of followers, and even losing fans.

3.6 Conclusion

This research studied the users popularity evolution in online social networks with a focus on professional users such as companies, celebrities, brands, and etc. To this end, the number of fans of almost 8K of the most popular professional users was collected in six daily snapshots, over a period of 14 months. The users' published posts were also collected in the same time period, which eventually provided around 20 million snapshots of popularity values. The experiments conducted on this data reveal interesting results. Users were categorized into two main groups fan-losers and fan-attractors, and four different patterns of popularity



(a) Cluster-1

(b) Cluster-2

(c) Cluster-3

(d) Cluster-4

Figure 3.7 – CDF (and BoxPlot) of number of published posts per user in the first (M1) and last (M14) months of the dataset (red dot in boxplot represents the Mean value of the distribution).

evolution were identified. Several factors are identified that influence the popularity trend of users, such as the social position like celebrities, external events associated to the owner of the page, and the level of activity. The findings from this study provide a comprehensive view on professional users' popularity evolution, and reveal the impact of different factors on it.

This study only analyzed professional Facebook users. The analysis of cross-popularity of these users on other major social networks, e.g. Twitter, Instagram, etc., can be considered as a future work. Beside the activity and external events, it could be very interesting to look on other potential influential factors such as specific strategies that users are following in social media. Providing a comprehensive list of suggestions for users to enhance their success in social media can also be an extension of this work.

Subpart B

Long-Term Evolution of User's Popularity

3.7 Abstract

As we observed in Section A, Facebook *Fanpages* show different popularity behaviors including fan-attracting and fan-losing. In this section, we aim to study the changes of user's popularity behavior in a period of four years. The main concerns are to find the variation pattern of popularity trends and identify the corresponding reasons. The study will investigate the effect of some influential factors on the popularity patterns of *Fanpages*. The results could help *Fanpages* to manage their popularity by being aware of the impact of their activities.

3.8 Introduction

The present study will extend the previous research on popularity evolution of Facebook *Fanpages* (Section A) by investigating long-term evolution of popularity patterns. As mentioned, *Fanpage* is a type of Facebook profile with some general attributes, such as the page description, events, posts and its category (selected by the page's owner). The main dynamic and valuable attribute on *Fanpage* is the number of fans, "Liked" number, precisely. We refer to the fans number of each *Fanpages* as their N_f in this thesis.

Regarding as a continuous study of our previous work [18], this subchapter aims to model the *Fanpages*' popularity evolution in a long-term duration and identify the general relationship between N_f and investigated data including category, posts and interaction level. To this aim, popularity data of Facebook *Fanpages* is collected in an extended duration additional to the first period in Section A. The difference of popularity patterns, identified in two separate periods, is compared and the characteristics of each pattern are studied. We also identify the main factors that affect popularity growing trends by investigating the impact of activity and external events. With the extended dataset, it is more persuasive to identify the variation of popularity trends and predict the tendency of a given pattern in future.

Our dataset includes nearly 8K *Fanpages* that have the highest number of fans validated by a third-party portal *Social Bakers*[4]. The data collecting work is specifically sophisticated and time-consuming because we were dealing stream data, not offline historical ones for four years.

3.9 Data Collection and Dataset

We follow our previous methodology [18] and start with a large list including 10K of top Facebook *FanPages* based on their N_f from the previously mentioned third-party application *Social Bakers* [5] which ranks users based on the number of fans.

Three crawlers are implemented to collect data as follows: Firstly, a data collection tool that queries FB public API[6] is used to collect the number of fans of users. The data

[4]http://www.socialbakers.com/

[5]www.socialbakers.com

[6]https://developers.facebook.com/tools/explorer

Table 3.3 – Dataset Characteristic

Attribute	**Value**	
	Phase1	**Phase2**
Crawling Period	Sep'13-Oct'14	Feb'16-Mar'17
Duration	14 months	14 months
#Users (#*FanPages*)	7,875	
Total #Samples in dataset	20M samples	18.8M samples
Avg(#Sample) per user	1,298	1,323
Avg(#Valid days) per user	327 days	345 days
Total #Post in dataset	43M posts	21M posts
Avg(#User_Posts)	1,865 posts	1,542 posts
Avg(#User_Post) per month	158 posts	152 posts

collection is performed for the selected 8K users over a period of 4 years discretely from September 2013 to March 2017. To have detailed information of popularity evolution, the value of N_f is recorded every 4 hours (6 times per day). We chose two separate 14-month durations to analyze the popularity pattern of facebook professional users. The reasons why we selected two separate 14-month durations are keeping coordination with our previous research and being able to study popularity trends variation in a long duration.

The second crawler collects user's category which determines the business sector of the page. Our selected 10K (8K after filtering) pages belong to hundred of categories. We pick up only 17 categories which contain enough *Fanpages* samples and fans number to map on the popularity trends and see which category is more popular among followers. The third crawler collects the activities (published posts) of *Fanpages* and their associated attributes in the mentioned periods. We record the comments, likes and views numbers of each interactive post, picture and video uploaded to the pages to gain persuasive information.

Based on our observations of users' popularity trends within a month which do not show a lot of peaks and drops, we chose month as the interval to report users popularity in. Thus, the monthly popularity vector is defined per user to represent the average values of user's N_f on each month. Moreover, we performed two refinement filtering over the collected data by the first crawler. In order to have a fair comparison, we excluded the users who do not have the accurate values of monthly averages because we missed some snapshots's data as the consequences of Facebook API limitation or network connection interruptions. The second step of refinement was to exclude the pages in the *Interest* category from the dataset.

A special category called *Interest*, based on what Facebook users are interested in. Most *Interest* pages display a Wikipedia article related to the page's name if one is available. These pages do not have any individual owner neither any published post, which are not fit to our requirement. After these three clawers and two filtering steps, we have a set of around 8K *FanPages* which build our dataset to perform this study. The importance and uniqueness of our dataset is because of the time-series data which has been collected continuously and live. A summary of the main characteristics of the dataset is presented

in Table 3.3.

3.10 Evolution of Popularity

The aim of this section is to analyze the evolution of users' popularity in overall. First we study the popularity trend for a set of special users namely *GlobalPage* and next we analyze the aggregated popularity evolution of all users available in the dataset.

3.10.1 Global Page

In our experiments, we identified a group of users who have a sudden and huge peak in their N_f in a very short period of time. With detailed investigation, we found that this peak reflects the impact of a service announced by Facebook, named *GlobalPage* [65]. Facebook *GlobalPage* is a special page structure for brands (e.g. large and international companies) which are well-known worldwide. This type of users used to have several separate pages with the same name but active in different languages and different locations (e.g. countries or continents).

If a brand has different pages and starts to utilize the *Global Page* service, which is not free of charge, there will be only one page representing that brand thereafter but will provide different types of information (language, posts, etc.) in different locations. Facebook shows an aggregated #fans in all of the pages that belong to the same brand and are merged in one global page. While it unifies the popularity metric (#fans) across all those pages, the content of the pages can still remain different [65]. This service provides an easy management for page manager and a better global visibility for brands.

To gain a better understanding of this phenomena, Table 3.4 provides a shortlist of brands in our dataset with the highest number of merged pages (#Pages) using this service. It's worth to mention that the number of merged pages for each brand can be more than the values in Table 3.4 because this data only refers to the number of pages that are available in our dataset. According to this table, each brand has several pages with different #fans in M1 when merging has not been done yet. Their #fans vary from 700K to 94 millions. While after merge, it has been enhanced to an accumulated big number for each brand in M14. It might possibly attract more people to follow them because of a raised and big number of fans.

The merged pages remain integrated until the end of the second phase, except *Nike*, which parted all of its integrated pages available in our dataset from its main page. Because of unstable variation of #fans as a result of using global page service, which is not presenting the effect of user's real followers and its natural popularity behavior, we excluded almost 10% of pages from our dataset in which they have been merged as global pages. It will be worth to explore the popularity behavior of global pages before and after merge in more details and investigate the effect of localized content and accumulated #fans on attracting new followers in a separate study.

Table 3.4 – Sample *Global Pages* with the highest number of merged pages (#Pages) available in our dataset. Columns third to sixth show the popularity of pages in 4 snapshots of data collection period. The values of #Fans are in Millions.

Brand	**#Pages**	**#Fans**			
		Phase1		**Phase2**	
		M1	**M14**	**M1**	**M14**
McDonalds	12	0.7 - 29.5	44.6	61.9	69.1
Facebook	10	2.2 - 94.9	164.2	169.8	184.7
KFC	10	0.7 - 6.4	35.2	39.9	44
Pepsi	9	0.7 - 17.6	33.4	34.7	36.2
Pizza Hut	9	1.1 - 10.7	15.7	27.2	29.1
Nike	8	0.8 - 4.3	38.1	42	0.001 - 42.5

3.10.2 Popularity Analysis - Genral

As mentioned earlier, the analysis presented in this study is based on monthly intervals of users' popularity. Monthly popularity is defined as the average of users' daily N_f in one month. Since our dataset covers 28 months in two separate phases, two 14-entry vectors are assigned to each user to represent their popularity trend within those two phases.

Figure 3.8 illustrates the cumulative distribution function (CDF) of users' popularity in four selected months including the first and last month of each data collection phases (P1-M1, P1-M14, P2-M1, and P2-M14) with the median values of 1.3M, 1.74M, 1.67M and 1.79M respectively. As it shown, the median value increased from M1 to M14 by 30% (and 38% increment for mean value - shown by red dot in boxplots) at phase1, then it is slightly declined to 1.67M at M1 of phase2 but raised again to 1.79M at P2-M14, with almost 50k more than its value at P1-M14. Although the popularity of users has substantially risen in P1-M14 compared to P1-M1 in the first phase, the distribution plots after P1-M14 and even in phase2 show that users' popularity values have remained almost unchanged in the overall view.

Exploring the variation of N_f from P1-M1 to P2-M14 for each user, despite its likely peaks and drops at this interval, shows that 70% of users have attracted new fans, called fan-attractors, and on the other hand, 30% of users have lost their fans, called fan-losers, during a period of almost 4 years. In a separate analysis for each phase, these numbers are 80% and 20% at phase1 and 54% and 46% at phase2 from the first month to the last month of the phases for fan-attractors and fan-losers respectively. It shows that users are mostly successful at phase1 in terms of attracting new fans.

Although the percentage of fan-losers at phase2 of the dataset shows that a noticeable number of pages are losing their fans during this period, the similar popularity distributions at P1-M14, P2-M1 and P2-M14 in Figure 3.8 indicate that the decline of N_f for fan-losers in the second phase might be very small amounts. To clarify the decline or raise amount of users' N_f, we compute the growth rate of N_f for all users in different intervals by subtracting the users' popularity on the first month from their popularity on the last month of each

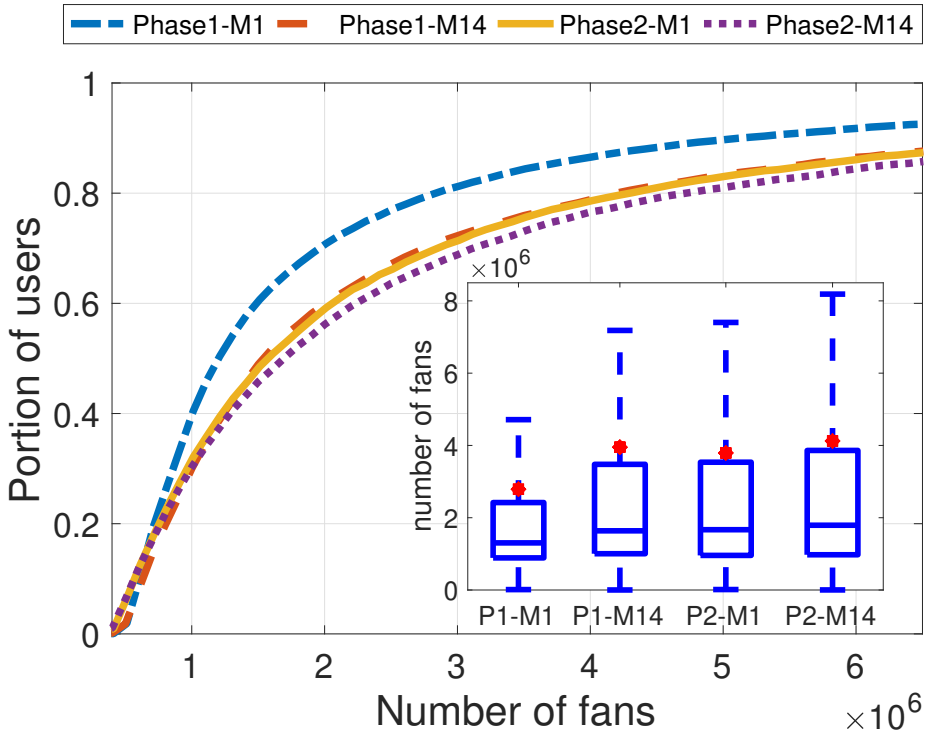


Figure 3.8 – users' CDF with the box-plots of #fans (N_f) distributions including two phases of M1 and M14. Red dots represent the mean values of N_f in particular month.

interval and show the growth distribution in Figure 3.9. Growth rate in Figure 3.9 presents the average raise or decline of users' N_f per month.

Starting from fan-losers in this figure whose growth rate is less than zero, the first phase has less number of them in compared to the second phase, with low percentage of decline in N_f. In contrast, the second phase has the higher number of fan-losers with the higher percentages of drops in N_f in compare to the first phase.

Regarding fan-attractors whose growth rate is bigger than zero in Figure 3.9, most of the users were successful to attract new fans and increase their popularity in the first phase. Their population value is decreased from the first phase (80%) to the second one (54%), indicating further lost or saturation in N_f. Following users' growth rate in different time spans illustrates that majority of users attract new fans at first phase but their population has been decreased gradually after M14. We will study the popularity stream of fan-losers and fan-attractors with more detail in section 3.11. To conclude the overall analysis of users popularity, most of the users attract new fans at first with higher percentages of growth rate but their popularity get gradually saturated in a later period (P1-M14, P2-M1 and P2-M14 in Figure 3.8) and their growth rate steadily converges to zero (Phase 2 in Figure 3.9).

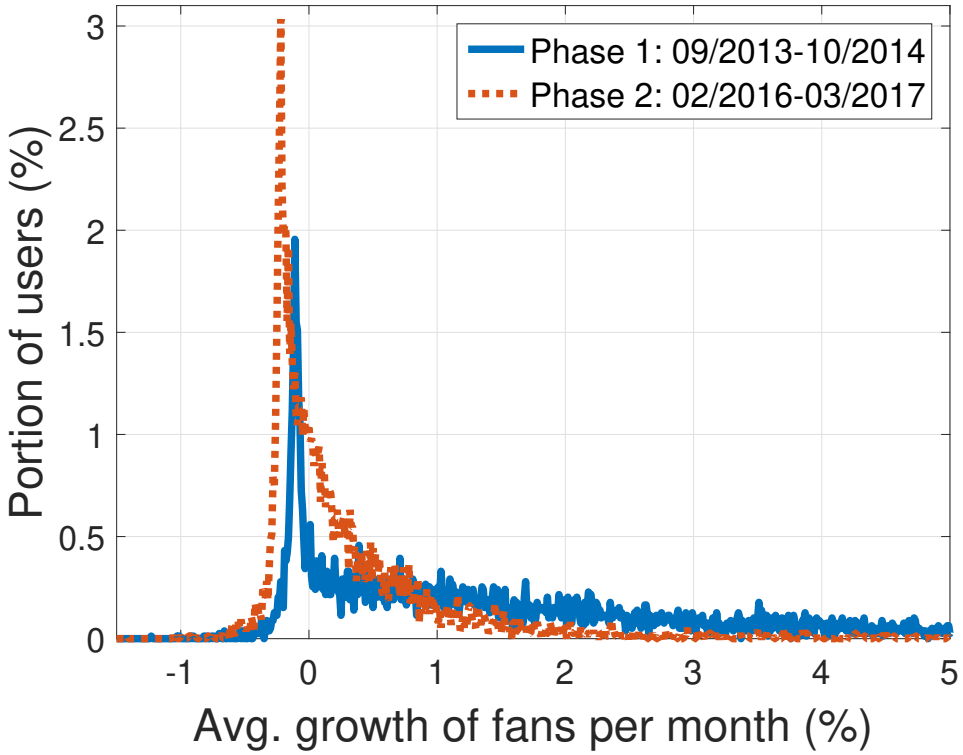


Figure 3.9 – Distribution of users' N_f growth rate in three time spans.

3.10.3 Popularity Analysis - Category Wise

In the time of subscribing a new *FanPage*, the page owner needs to select a category for the page from a pre-defined set of categories provided by Facebook. This category specifies the business sector of the page. The selected category is available in the profile of the user and our second crawler collects this information.

In this part, we chose 17 (out of 158) most populated categories whose number of pages cover more than 75% of the users in the dataset. Table 3.5 represents a brief summary of the population and the growth of N_f for each category in the two mentioned periods. The main observations from this table are as follows:

i. The fifth and sixth columns of Table 3.5 refer to the percentage of the average growth of users' N_f in each category. Reduction of growth rate from phase1 to phase2 observed in tracking of overall popularity in the previous section, is observable for each category as well. There is no category that has over 50% of average growth at phase2 but six categories at phase1. It seems that the most of *fanpages* have reached the saturated point of their number of fans during phase2. Since these pages have already grown to a huge number of fans and most of the fans have already followed them, it is normal to not have an explosive enhancement in N_f later on.

ii. Comparing the percentage of average growth shows that only the *Product Service*

Table 3.5 – Populated categories distribution in the dataset. Fifth and sixth columns represent the average growth rate of N_f over the two periods of collected data.

#	FB Category	#Page	%Pages	%Avg. growth	
				Phase1	**Phase2**
1	Musician Band	1231	**17**	47	1.2
2	Community	986	13.7	**2.1**	**-1.8**
3	Tv Show	477	6.6	53	2.5
4	Movie	413	5.7	28	4.9
5	Food Beverages	302	4.2	19	10.6
6	Product Service	267	3.7	**24**	**37.6**
7	Public figure	246	3.4	64	8.9
8	Company	188	2.6	23	13.9
9	Athlete	188	2.6	**101**	6.1
10	Actor Director	179	2.5	**97**	5.2
11	Entertainment	166	2.3	26	7
12	App page	143	2.0	17	-0.2
13	Clothing	139	1.9	29	8.5
14	Media News	134	1.8	76	15.4
15	Sports Team	125	1.7	**92**	9.2
16	Games Toys	109	1.5	13	0.7
17	Health Beauty	85	1.2	17	17.7

and *Health Beauty* categories have bigger growth in phase2 than phase1. Others, like category *Athlete, Actor Director, Sports Teams* and *Media News* and so on, have mostly huge drops in their average growth from phase1 to phase2. The same explanation about the saturation mentioned above can justify those drops in categories as well.

3.10.4 Popularity Analysis - Activity Wise

Publishing posts is the main activity on social media that professional users take advantage of it to promote their products, news, events or any other relevant thing. Furthermore, it enhances the prospect of attracting new followers [21]. In this subsection, we aim to investigate whether publishing posts effects number of followers of FanPages.

Figure 3.10 shows the CDF of the number of posts which published by the pages in the four selected months of our dataset. According to the box-plots, the mean values of the number of posts (shown by red dots) have a very slight decrease from the P1-M1 to P2-M1 by almost 10% drop. However, the mean value turned out to be over 160 posts at P2-M14. As shown in the previous section, the growth rate of users' popularity is also declined during the second phase of the dataset (Figure 3.9). According to the CDF curves in Figure 3.10, the variation of the number of published posts from phase1 to phase2 is too small to cause this decline of the popularity growth. It conveys that even though users keep

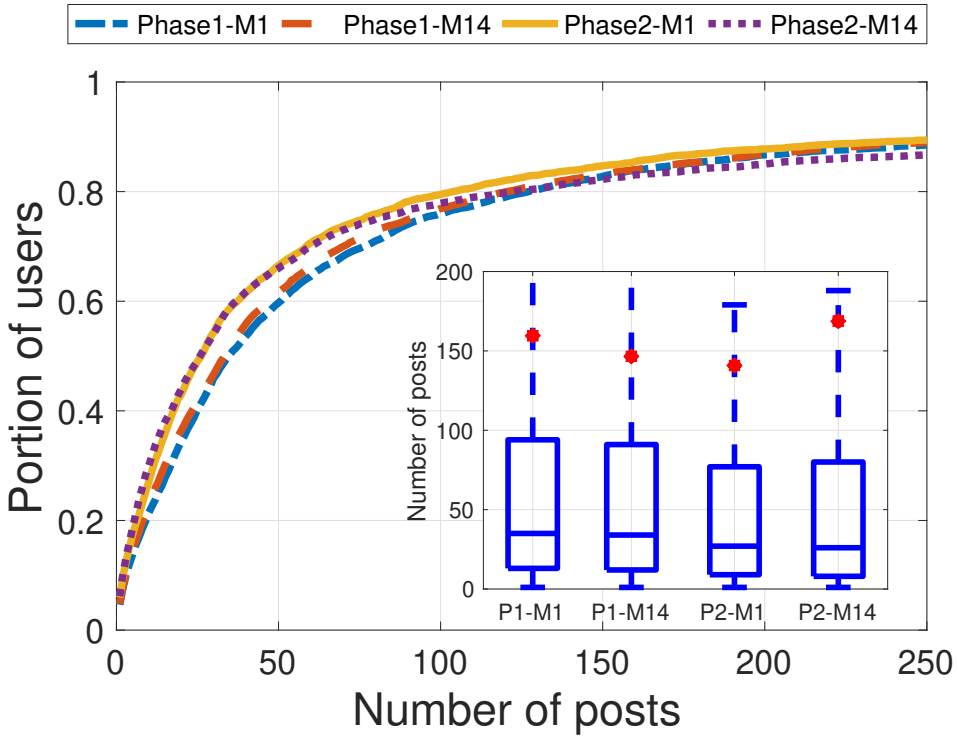


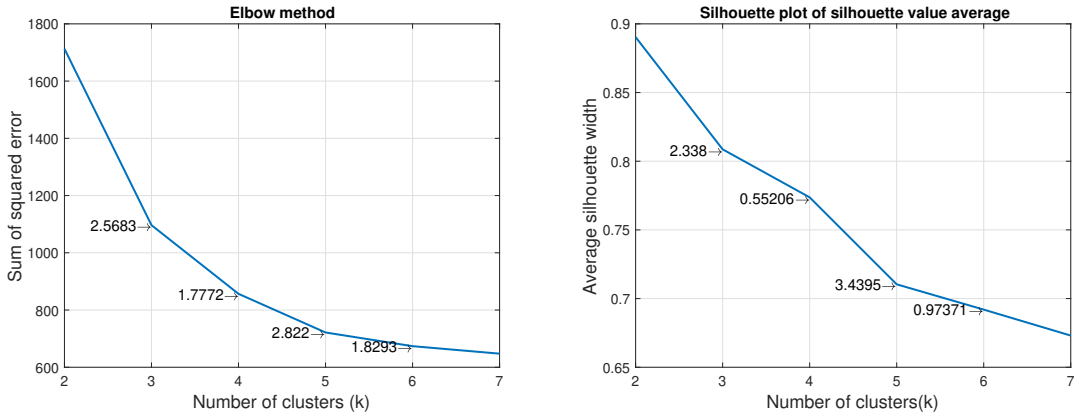
Figure 3.10 – CDF of the number of posts published by users in the first and last months of phase1 and phase2. Red dots represent the mean values.

to publish almost the same number of posts during the 4 years of our study, the popularity of their pages has been almost saturated during phase2 and it has not grown as fast as we observed during the first phase.

In a nutshell, this section provides some simple findings on the overview popularity evolution in our dataset as well as some interesting observations in different categories. In the next section, we explore the popularity evolution of users in detail and identify different clusters of their popularity trends.

3.11 Users Clustering Evolution and Comparison

This section aims to identify different clusters of users with similar patterns in their popularity evolution trends and compare those patterns in phase1 and phase2. As for the previous study [18], we aim to analyze the variation of users popularity trends from the first phase to the second phase of the dataset using proper clustering techniques. We will provide a deep analysis of correlation between popularity trends and N_f in *Fanpages* and categories. At the end, different characteristics such as popularity level and activity distributions per each identified clusters have been investigated to conclude the evolution of N_f



(a) Sum of squared error (SSE)

(b) Silhouette width for different cluster numbers k

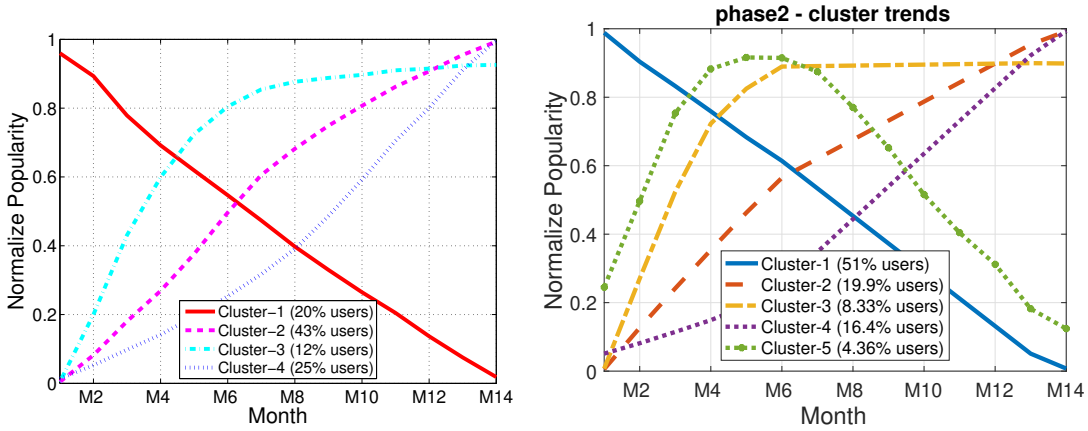
Figure 3.11 – SSE test to find the appropriate k value for our database

model.

3.11.1 Extended Feature Vector and Clustering

To cluster professional users based on the popularity values, two 14-month popularity vectors of each user are used as the feature vectors in the clustering methods. The feature vectors thus represent the time-series popularity trends of users where different patterns of fan variations can be identified by clustering. The goal is to identify and group the users with similar popularity evolution behavior into a single cluster, regardless of the value of N_f. For example, consider two *FanPages* from quite different ranges of popularity, which both have 50% growth of N_f with the same trend over the same time period. They should be grouped into a same cluster because their popularity trends are very similar. To this end, the feature vectors should be normalized to bring the entries into a comparable range. We used the Min-Max normalization method which scales every entry of feature vector into $[0, 1]$ by obtaining the values 0 and 1 at the minimum and maximum points of each vector, respectively.

As for the previous study [18], we identified four different popularity trends in the first phase of the dataset shown in Figure 3.12a. In order to compare the users' behavior in phase2 with phase1, the same clustering method is applied to the 14-month feature vectors in the second phase as it is done for the first phase in the previous work. Elbow method and average silhouette width [70] have been used to decide the optimal number of clusters at phase2. Figure 3.11a and 3.11b shows sum of the squared errors (SSE) for different values of number of clusters (k). The elbow point determines the optimum number of clusters in both figures. In order to assure the elbow point, we calculate the second derivative on each k point in Figure 3.11a and 3.11b to see the maximum slope changing rate, which represents the optimum point. As shown, points at k equals to 5 are the maximum of all the points, 2.822 and 3.4395 respectively. To sum up, k equals to 5 can be an appropriate



(a) Normalized popularity trends (k=4) (b) Normalized popularity trends (k=5)

Figure 3.12 – Popularity trends in two phases

number of clusters in the phase2, regardless the phase1's cluster number which was 4.

Figure 3.12a and 3.12b represent the normalized popularity trends for the identified clusters in both phases. Each plot shows the average value of the normalized N_f belonging to the users in the corresponding cluster. In general, both phases show three different ascending patterns, which are named as cluster-2, cluster-3 and cluster-4. The popularity trend of Cluster-1 shows a monotonically descending behavior. In phase2, there is a collection of *FanPages* grouped in cluster-5 that their average popularity reached a peak between M5 and M6 and then descended into its initial value at the beginning of this period. We compare the result in phase2 3.13b with phase1 in our previous work 3.13a and we infer the following points:

i. In phase2, there exists an extra popularity evolution pattern (cluster-5) with only 4% of users that shows a trend with the maximum value between M5 and M6. After M6, the fanpages in this cluster start losing their fans quickly. The descending rate is even slightly greater than cluster-1. We will have a detailed investigation on those pages to find out the reason behind this behavior.

ii. Considering cluster-1 in figures 3.12a and 3.12b, there are now over 50% of users in cluster-1 of phase2, 30% more than phase-1, who are losing their fans. It means that the ascending popularity pattern of about 30% of users, who used to attract fans in phase1, now turned into a steadily descending pattern in phase2.

iii. In phase1, the most populated cluster was cluster-2 by around 43% of users. This cluster's population dropped to 20% of total users in the second phase. Since the popularity behavior of this cluster is continuously incrementing and the fanpages in this cluster were the most successful ones in attracting new followers at phase1, this drop can be interpreted as an almost 20% drop in the most fan-attracting pages from phase1 to phase2.

iv. Cluster-3 in phase1 was the least populated cluster where its number of pages dropped about 25% and now it contains almost 8% of fanpages at the second phase. As its pattern shows a saturation of popularity, exploring the users who moved from this cluster (saturated popularity) at phase1 to the others (increasing/ decreasing popularity) at phase2 will be interesting what we will discuss it in the following subsections. We will also investigate the population drop which happened in Cluster-4.

Before we go through the clusters' detailed analysis, here comes three main questions related to the identified trends. Firstly, it is important to know what percentage of fanpages from ascending behavior clusters (including clusters 2, 3 and 4) at phase1 have been moved into cluster-1 at phase2. In other words, we aim to identify fan-loser pages and to understand whether those fanpages are related to some particular categories and business sections. Secondly, does there exist any exceptional fanpages which their descending trend is turned to an ascending one? More precisely, we are interested to identify those fanpages who jumped from cluster-1 at phase1 to any other ascending clusters at phase2. Finally, despite the most fanpages that are now losing their fans during phase2, some of them still maintain their ascending trend, which are interesting cases for us to identify and explore their activities.

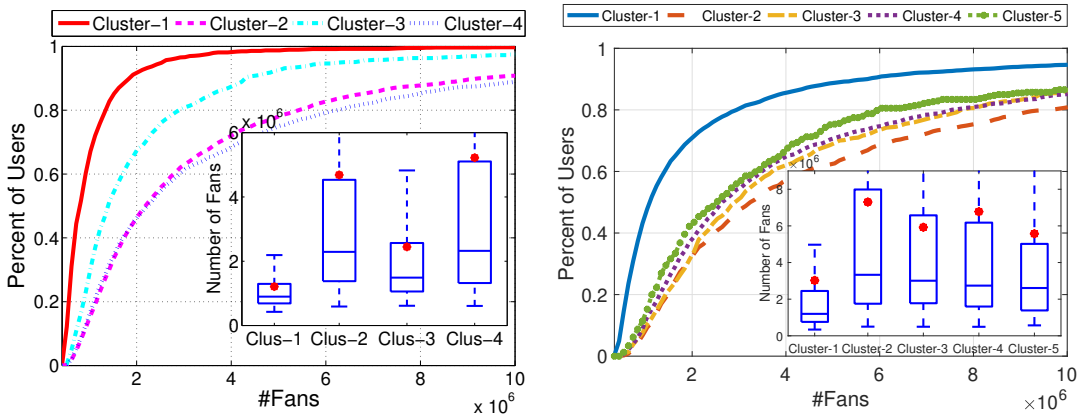
Next, we discuss about the mutual distribution of clusters' population in the first and second phases. The detailed information of fanpages, such as category and activity, will also be investigated with respect to their clusters. The analysis results will be compared with the previous study.

3.11.2 Popularity Distribution of Clusters

In this section, we analyze the CDF of N_f of users for each cluster in phase1 and phase2 and compare the clustering results with regard to the users' number of fans during those two phases. The aim of this analysis is to identify the relation between the absolute value of N_f and the normalized trends. Figure 3.13 shows the CDF of N_f in two phases.

First, according to the box-plots in this figure, the average of users' N_f in cluster-1 is grown about 2M from phase1 to phase2. It illustrates that the enhanced population of this cluster is related to some users with high number of fans which are moved to this cluster due to losing their fans in the second phase, while they were in three other fan-attracting clusters in the first phase.

Second, among three ascending trends, the mean values of cluster-2 and cluster-3 show a big growth from phase1 to phase2. According to the CDF plots, N_f of users who are grouped in those two clusters in phase2 are much higher than in phase1. On the one side, the increase of the mean values and CDF distribution indicate the high popularity growth of users of those two clusters. On the other side, since there is a drop in those clusters' population, we can also judge that the users of those two clusters in phase1 with less N_f are moved to other clusters in phase2 and caused a great growth in the mean values of users' popularity. Third, among three ascending trends, the CDF of cluster-4 at phase2 remains almost same as its CDF at phase1. As its population is not significantly changed,



(a) Distribution of users' N_f in four identified clusters in phase1. [18]

(b) Distribution of users' N_f in five identified clusters in phase2.

Figure 3.13 – CDF (and boxplot) of the distribution of users' N_f at phase1 & phase2. Red dots represent the mean values.

its popularity CDF variation can be because of regrouping less popular users of this cluster into other clusters, and more popular users to this cluster after the second clustering.

Finally, cluster-5 with only 4% of *Fanpages*, represents a similar popularity distribution to other fan-attractor clusters, where the fans evolution pattern are divided them into separate clusters. Based on this observation, we can assume that cluster-5 may consist of those fans-attractors.

3.11.3 Users' Long-term Popularty Behavior Variation

This section aims to explore the variation of users' group behavior from phase1 to phase2. Figure 3.14 shows the migration of users from the clusters in phase1 (along the X-axis) to the clusters in phase2 (along the Y-axis). We will investigate the population of users whose popularity trend has been changed from phase1 to phase2. This change of popularity behavior is recognizable when a user get reclustered at phase2 in a different cluster than its cluster at phase1. Looking at Figure 3.14 illustrates that the majority of users in cluster-1 at phase1 have again grouped in cluster-1 at phase2. It reveals that fan-loser users (representing descending popularity trend in cluster-1) in phase1 keep losing their fans during the second phase as well. However, there are few professional users that succeeded to jump from cluster-1 at phase1 to other clusters at phase-2 who might be good examples of improving their fan attraction policy.

Considering cluster-2 at phase1, the interesting point is turning the popularity trend of 14% of fanpages out of 43% (the population of Cluster-2 at phase1) from ascending to descending behavior. Regarding cluster-3 at phase1, there are almost 60% of *Fanpages* go to cluster-1 at phase2. We can say that if *Fanpage*'s popularity trend acts almost saturated in the end of the period (end of the first phase), there is 60% possibility that its trend will

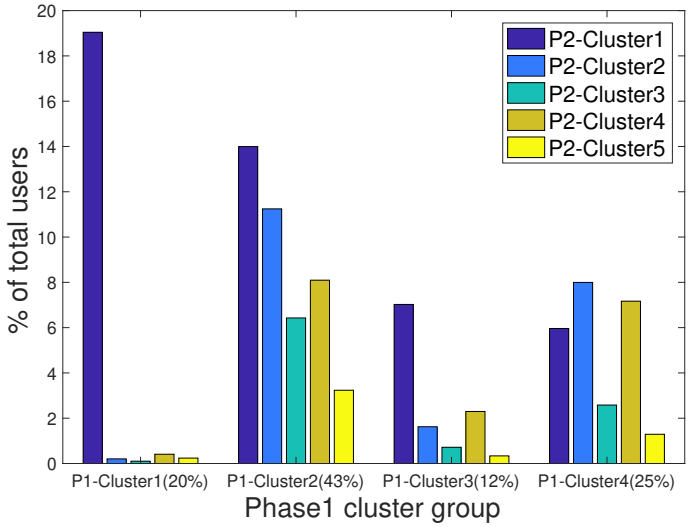


Figure 3.14 – Cluster group percentage from phase1 changing to phase2

start to decrease. The last one is cluster-4, and it shows no particular cluster for changing tendency. In spite of no dominating cluster, there are over 60% of total *Fanpages* stayed in fan-attractor clusters (cluster-2 and cluster-4).

In the later discussion, we will focus on the category details of those percentage mentioned above. We will investigate the categories of *Fanpages* that are now popular compared with the time in phase1 and phase 2. Furthermore, we will look at the categories used to be popular but now they are losing their reputation and fans. The influence of any possible events that turned their trend will be studied. We will try to find the answers and reasons to explain the mentioned variations.

3.11.4 Category Analysis

In this part, we investigate the distribution of categories inside the identified clusters to understand if there are categories with a dominant population in a specific cluster. Moreover, we analyze the combination of the phase2 clusters consisting of the phase1 clusters so as to see the details of variation between clusters in two phases. We aim to analyze the dominated categories in phase2 regarding all clusters and also compare the difference between phase1 and phase2.

Figure 3.15 shows the distribution of the 17 most populated categories, mentioned earlier in Table 3.5, across the five identified clusters in phase2. Figure 3.16 shows the phase2 cluster combination from phase1 cluster in category-wise mode. We apply stack bars to represent different percentage of phase1 clusters, and consider a group of whole stack bars in each category to be phase2 clusters, ordering from left to right as cluster-1 to cluster-5.

First, according to Figure 3.15, there are five high presence of categories in cluster-1,

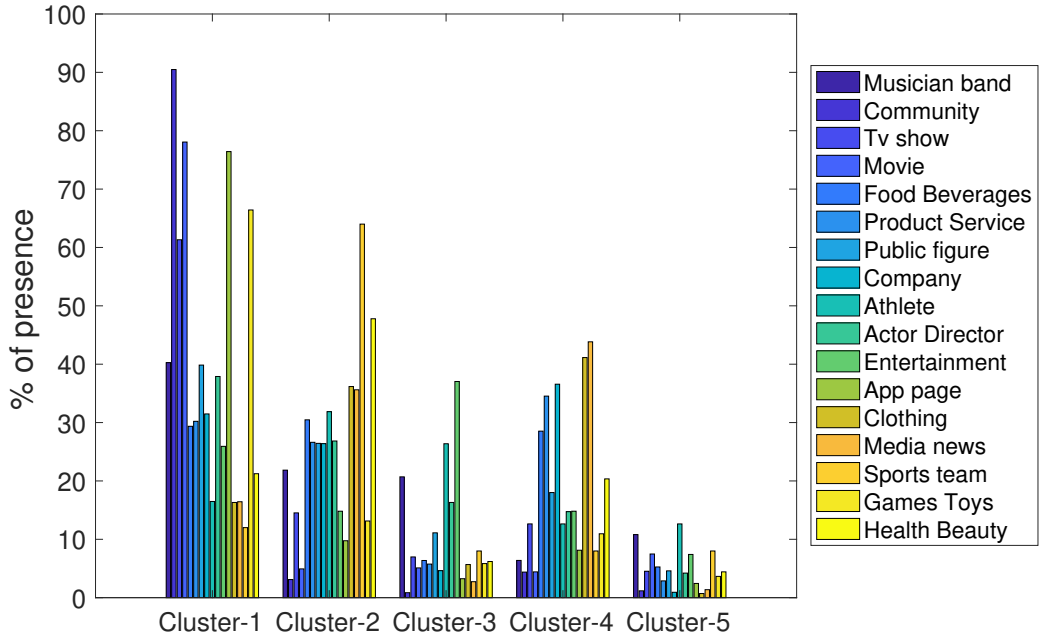


Figure 3.15 – Predefined Facebook categories distribution in each identified cluster.

such as *Community*, *Movie*, *App Pages*, *Game Toys* and *TVshow*, with over 60% portions in each category respectively. A *Community* pages even have higher than 90% of them in cluster-1, considering as the almost no change from phase1 to phase2 in Figure 3.16. Given that the users in this cluster are losing their fans, and the *Community* category is the second most populated category with 13.7% of the users in the dataset, it can be concluded that it is also the biggest set of fan-loser users.

According to Facebook[7], "a Community Page is a page about an organization, celebrity or topic that it does not officially represent. A Community Page has a label below its name that identifies it as a Community Page and links to the official page about that topic." Our observations show that a *Community* page is a place that Facebook users gather to share their ideas, images, posts around a specific topic, company, or celebrity and cannot remain attractive to users over time. Although these pages are in our dataset because of their high popularity (N_f), but according to the clustering results and the category distribution, they have lost a portion of their users over the period of this study. One of the reason we found is the new feature of Facebook "Verified" which provide the possibility for verifying popular pages which Facebook started in May 2013. After verification, people are more likely following the verified pages instead of the community pages.

Others pages like *Movie*, *App Pages*, and *TVshow* are used to be fans-attractors. Now in phase2, a majority of them jumped from other phase1 cluster to phase2 cluster-1. In

[7]https://www.facebook.com/help/187301611320854/

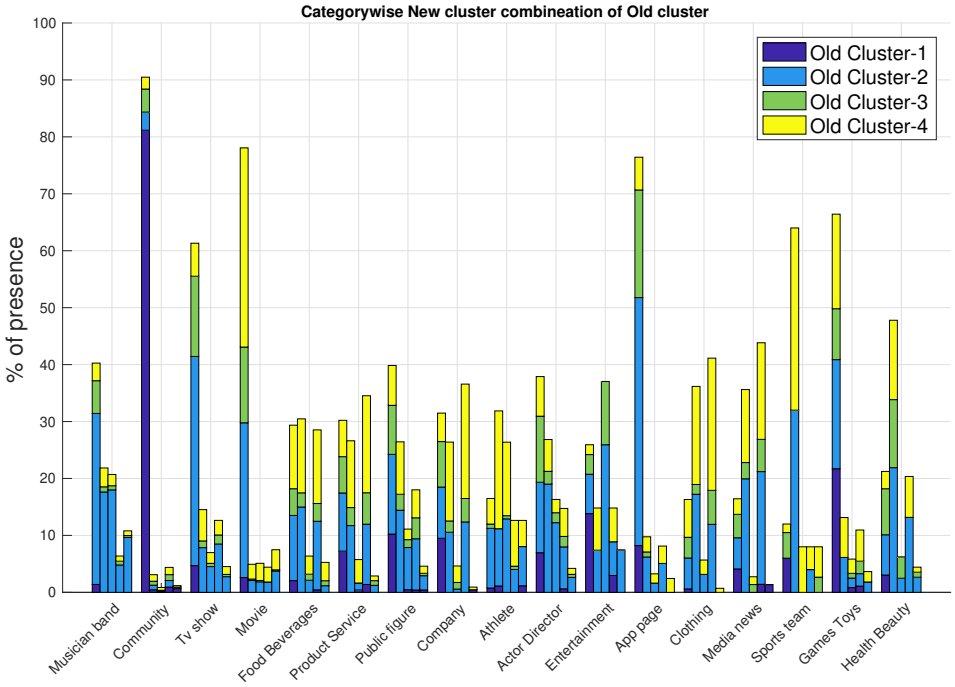


Figure 3.16 – The combination of phase1 cluster group in phase2 cluster in predefined Facebook categories distribution.

summary, according to the popularity trend in Fig.3.12b and category distributions of cluster-1 in Fig. 3.15, we can infer that except those five categories of *Fanpages* mentioned above, the bigger portion of the other categories are still attracting fans although more than 50% of total professional users in our dataset are fans-losers.

Cluster-2, which shows a fixed rate of popularity growth, includes a high presence of *Sport team* and *Health Beauty* categories with 64% and 48% of users respectively in their own category. However, for *Musician band*, *TV show*, *Actor Director* and *App page*, those categories that used to be more than 50% of users in phase1 cluster-2 and now most of them change their cluster in phase2, especially *Musician band*, *TV show*. They are two of the main popular *Fanpages*, with 17% and 6.6% of users in our dataset. These two categories, accompanied by *Actor director*, contain most of the celebrities' pages in our dataset. As we look at the Figure 3.16, category *Musician band* and *Actor Director* separated their phase1 cluster-2 into each phase2 clusters, but more than 35% of *Fanpages* in category *TV show* changed from phase1 cluster-2 to phase2 cluster-1. We can conclude that most of the celebrities' pages are no more attract fans in a fixed rate in phase2.

The distribution of categories in *Cluster-3* shows only category *Entertainment* with over 30% of users and most of them came from phase1 cluster-2 and cluster-3. However, regarding *Athlete* having dominated minimum presence in phase1, there are over 25% of enhancement in phase2. All of them jumped from phase1 cluster-2 and cluster-4 to phase2

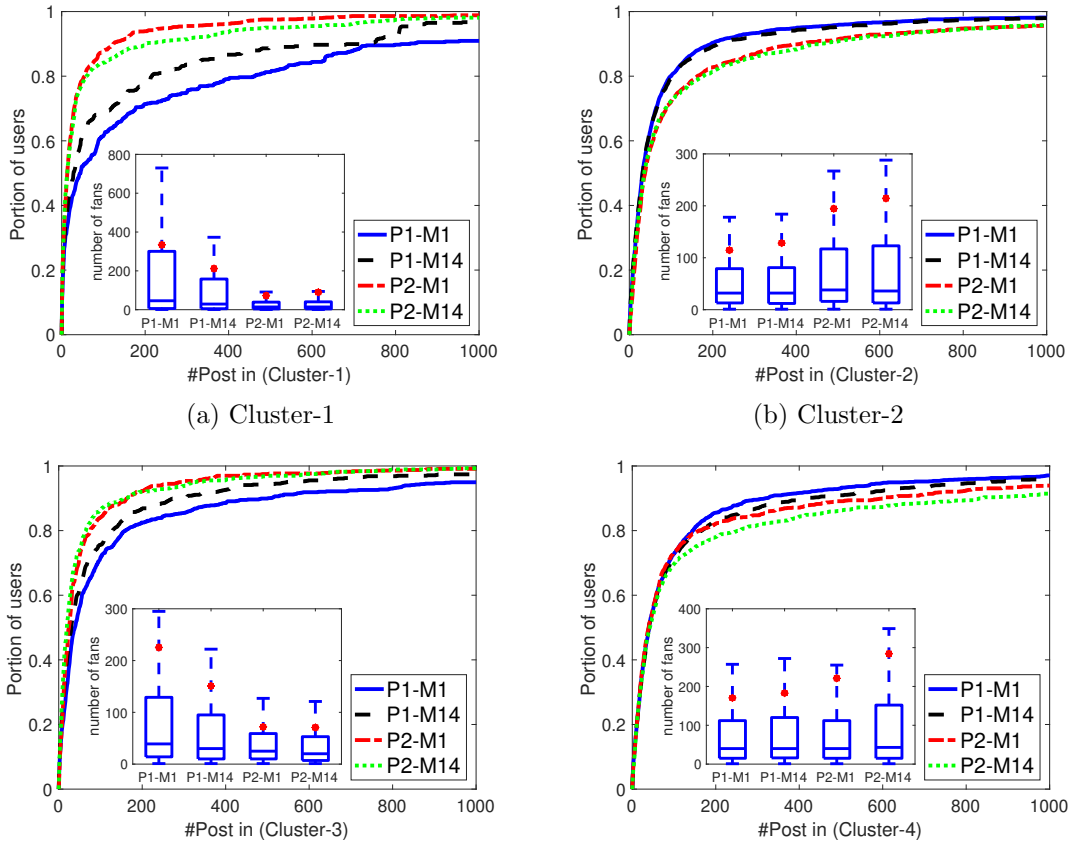
cluster-3, which represent those *Fanpages* turned to be saturated in the end of phase2. When we take a closer look about how phase1 cluster-3 change to phase2 clusters, we figure out that most saturating trend turned out to be descending trend. For example, we can compare the portion of phase1 cluster-3 separated in different phase2 clusters. especially categories *App page*, *TV show*, *Actor Director* and *Movie*. To sum up, we predict that most of users in cluster-3 cannot remain to be saturated all the time, and it has high possibility to turn its trend to be descending. This could have different explanations like fan-saturation, reduction of the activity or external events. In the next section, we look for in detail to understand the effect of activity volume on users' fan-trends as a probable influential factor.

Cluster-4, which includes 16.4% of our users, has a variety of categories distribution. Two categories, *Athlete* and *Sport team*, used to have more than 50% of their population in this cluster. However, they all drop lower than 20% in phase2. According to Figure 3.16, most of the users in those two categories changed to phase2 cluster-2, which means they still attract fans but in a fixed rate. In addition, Other categories, like *Clothing* and *Media news*, show over 40% of the *Fanpages* in phase2 cluster-4. Due to their combination in Figure 3.16, the phase2 cluster-4 is mainly consist of phase1 cluster-2 and cluster-4. Therefore, we can say most of the *Fanpages* belonging to *Clothing* and *Media news* remain attracting new fans in both phases we consider.

Because of the similar field between *Athlete* and *Sport team*, we investigate some famous celebrities and athletes, such as *Neymar* (Football player), *Real Madrid C.F.* (Sport team), in this cluster. For users such as those related to football, the most probable reason of significant N_f growth may be the main events of European leagues. These start on the middle of summer and end in the middle of the following spring. We will explain more about this phenomenon in Section 3.12.2.

Cluster-5 at phase2, which has no clue in the previous work, is clustered in phase2 because a non-neglectable percentage of users exist. Two categories *Musician band* and *Athlete* have over 10% of *Fanpages* in this cluster. Based on their trend in phase1, both of them are mainly fan-attractors and celebrities in public. It seems reasonable that a little portion of them turned their increasing rate into decreasing coincidently in phase2. We are curious about what happened to them during the second period and we will try to figure out in the next section.

To sum up this part, we firstly observed that *Community* is characterized as the most fan-losing category with a major presence in *Cluster-1*. Second, The categories relating to sport and health&beauty have the ability to keep attracting fans, with a significant presence in the two most fan-attractor Cluster-2 and Cluster-4, regardless in phase1 or phase2. Third, we consider cluster-3 to be a transition trend because most of users in cluster-3 easily changed to the other clusters in new period. Finally, we want to find the relative news or reasons why a little percentage of *Fanpages*, especially categories *Musician band* and *Athlete*, show a maximum peak in the middle month of phase2.



(a) Cluster-1 (b) Cluster-2

(c) Cluster-3 (d) Cluster-4

Figure 3.17 – CDF (and BoxPlot) of published posts number for first and last months per user in two phases. Red dots represent the mean values.

3.11.5 Activity Analysis

Being active in Facebook by continuously publishing new posts can ensure professional users to stay in touch with their followers and keep them updated with their relevant news or events. This sort of engagement in social media can affect the number of fans and previous studies shows that by actively publishing posts can also attract new followers [21].

To reveal the impact of activity on popularity, Figure 3.17 shows the CDF of published posts' number in two phases, each from M1 and M14 with four clusters. In Figures 3.17a and 3.17c, it illustrates that the published posts of the users cluster-1, who lost their fans, declined from the M1 to M14 in phase1 (by reduction of 12 posts in average). This situation can be observed for the distribution of users in cluster-3 in phase1 as well. As discussed before, the N_f of users in this cluster is almost constant for the second half of the study period. It can be concluded that the reduced number of activity in these two clusters is an important factor for the lost of fans in cluster-1 and the failure to attract new ones in cluster-3. Considering about phase2, it shows, however, a slightly increase in cluster-1

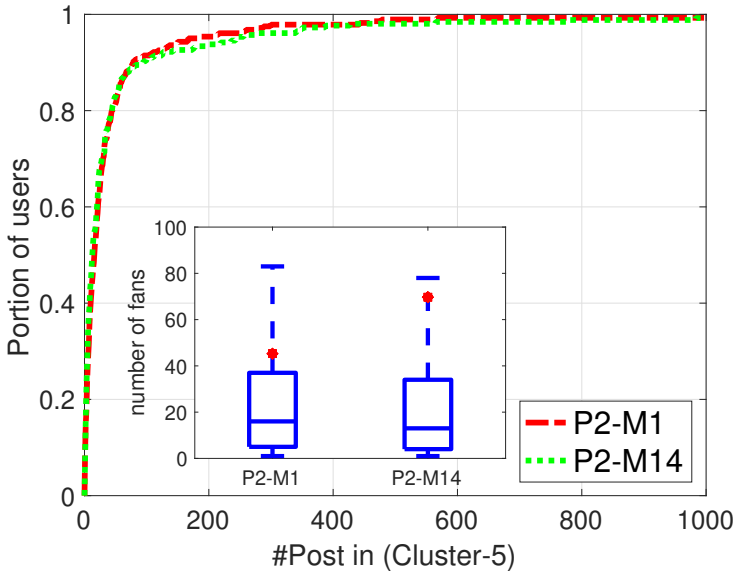


Figure 3.18 – CDF of number of published post for users in cluster-5

and unchanged posts' number in cluster-3. That is because we re-cluster those professional users again and the mix trend in second phase are balanced by the users changing from other clusters.

In contrast, the activity level of users remain steady in the two most fan-attracting clusters, cluster-2 and cluster-4, in Figures 3.17b and 3.17d. In cluster-4, we can even see a small increment in active posts number of users, who published more than average post numbers in the last month is greater than the number of users who posted that much in the first month. Considering their popularity trends which show a continuous growth, it can be deducted that being constantly active effect the process of attracting new fans.

The most important point in Figure 3.18 is a maximum peak of fans number in the middle of phase. Those users have the least average posts number of all clusters. It means that reducing active posts can easily turn the ascending trend into descending in a short time.

To conclude, we observed that staying active in terms of publishing posts can help to attract new fans and followers whereas reducing the activity level can lead to stagnant number of followers, and even losing fans.

3.12 Influential Factors on Popularity Trends

In previous section, we studied the popularity evolution of users and identified various characteristics of them in cluster level. In this part, we aim to identify the influential factors that affect the trends of popularity in user level and study the impact of two main factors, activity and external events of users.

Table 3.6 – Few samples of inactive users in phase1 and comparison in phase2. (*Monthly Growth)

Page Name	Page Category	Avg. M.G.* Phase1		Total Growth	Avg. M.G.* in Phase2		Total Growth	Type
		Absolute N_f	Growth rate		Absolute N_f	Growth rate		
Adele	Musician	1,051,471	2.8%	28%	132,861	0.32%	2.6%	attractor-attractor
Michelle Obama	Politician	257,930	2.5%	33%	184,670	0.91%	16%	attractor-attractor
Aamir Khan	Actor	101,866	8.2%	170%	-6,993	-0.08%	-0.6%	attractor-loser
Allen Iverson	Athlete	86,702	3.9%	63%	25,837	1.1%	10.6%	attractor-attractor
Robert Pattinson	Community	-3,420	-0.2%	-2.1%	-4,357	-0.24%	-3%	loser-loser
Zynga RewardVille	Games Toys	-10,825	-0.13%	-1.7 %	-16,755	-0.22%	-2.8%	loser-loser

Table 3.7 – Few samples of inactive users in phase2 and comparison in phase1. (*Monthly Growth)

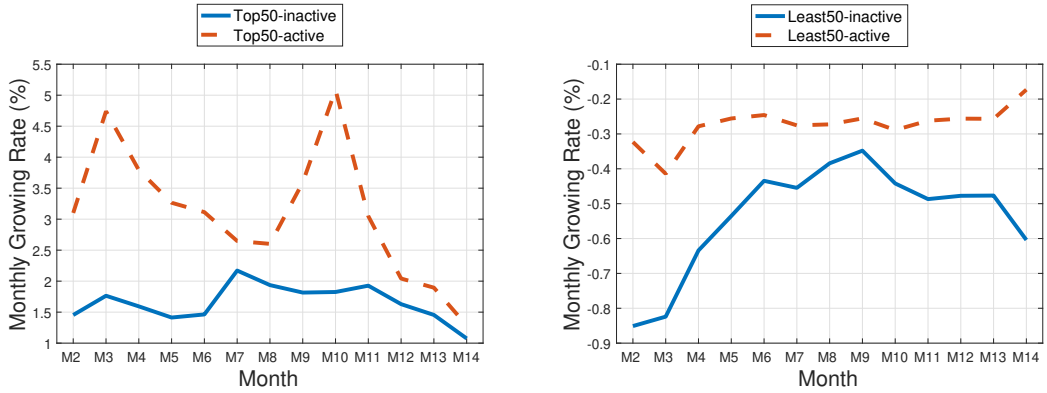
Page Name	Page Category	Avg. M.G.* Phase1		Total Growth	Avg. M.G.* in Phase2		Total Growth	Type
		Absolute N_f	Growth rate		Absolute N_f	Growth rate		
VICE	Media news	1,819,256	6.4%	124%	1,470,697	2%	30%	Attractor
Resident Evil	Movie	92,006	0.21%	2.9%	7,719,387	6%	258%	Attractor
Neuva Delhi	City	317,330	2%	29%	240,021	1.2%	15%	Attractor
Clarins	Health Beauty	437,830	2.3%	34%	207,150	0.84%	10.3%	Attractor
Alacakaranl	Book	-23,011	-0.11%	-1.4%	-80,454	-0.5%	-6.3%	Fan-loser
William Shakespeare	Author	-25,911	-0.06%	-0.8 %	-71,822	-0.2%	-2.5%	Fan-loser

3.12.1 Impact of Activity

Activity (publishing posts) in social media provides a golden opportunity for professional users to interact with their fans and also increase their visibility to attract new followers. As previously presented results in Section 3.11.5 show, users with high and permanent level of activity are among the most popular users in our dataset whose popularity is increasing as well. On the contrary, users with low active level, which means seldom publishing posts and low renewing rate, lost their fans and got less popular. This somehow indicates the positive correlation of activity and popularity which means that activity has a high influence on attracting new followers.

However, we identified a subset of popular users in our dataset who were not active at all and didn't publish in the period of our study, but surprisingly their popularity was still growing. The population of all inactive but still growing *Fanpages* are 10% (808 Users) of the dataset in the first phase and 2.5% (201 Users) of the dataset in the second phase. Here we define 'inactive' that *Fanpages* have no published posts, no sharing links and no renewing information during the defined periodLooking to the nature of these accounts shows that most of the users that are attracting new fans without any activity are very famous celebrities and well-known brands around the world. These inactive pages are attracting new followers because people usually add those pages to their profile's interest list in order to have their names in their profiles. On the other hand, the *Fanpages* in the other group of inactive users which are losing fan mostly belong to the personal web pages of people who are not very famous, or pages related to old-fashion. It seems that people follow them because of their temporary activities so that being inactive after events leads their pages to lose fans.

One of the different probable reasons behind this inactive behavior is the tendency of users to being active on other OSNs. Our manual inspection of few users (e.g. *Adele*, *Allen Iverson*) revealed that they are active mainly on Instagram or Twitter during the



(a) The top 50 users in active and inactive groups

(b) The least 50 users in active and inactive groups

Figure 3.19 – The monthly average growing rate for two phases

first period of being inactive on Facebook. And we also discover most of the celebrities and brands renew their *Fanpages* and start posting or sharing posts in the second period. We infer from this phenomenon that is probably because the function release during 2014 that professional users can link their *Fanpages* with OSNs, like Instagram and twitter for example. Considering that Facebook is still the main group of followers, they rearrange their pages and update their posts on either Facebook *Fanpages* and other OSNs.

We select few sample users from the subset of inactive users in phase1 and phase2 respectively. Table 3.6 and 3.7 includes totally 6 inactive users which has been chosen to focus on particular phases. The 12 chosen pages have a great positive (the first four users) or negative (last two users) monthly growth during the period. According to these two tables, for example, *Adele*'s page has the highest absolute popularity growth with more than 1M new fans in phase1, but these number dropped rapidly in phase2. Page *Aamir Khan* is another sample that his page's number of fans is increased by 170% in the first phase but decreased 0.6% in the second phase. *Resident Evil*'s page has about 258% huge growth of fans in phase2 but it shows only little growth when it was in phase1.

Now the interesting question is how the trend of popularity evolution changes for inactive celebrities in compare to active ones. Although, inactive celebrities are still attracting new followers in the period of their inactivity, we found that their popularity monthly growth is not as high as active celebrities' growth in the same period. In Figure 3.19 have proven this perspective. We collected both phases of active and inactive users based on their posts, activities and sharing numbers. Then we chose the top 50 and the least 50 users to calculate their monthly growing rate in both active and inactive data collection. Figure 3.19a represents the average of two phases growing rate per month, so does Figure 3.19b. We can easily come to the conclusion in Figure 3.19a that top 50 active users can always have higher monthly growing rate than inactive ones by 2% in average. The situation happen to Figure 3.19b that active users have average 0.3% less losing rate compared with inactive users.

For better clarification, we compare two celebrities from the same category but different activity behavior. We identified *Shakira* as an active user, who has published 23 new posts per month in average during phase1, in our dataset. This is a good sample to be compared with *Adele* as a sample of inactive musician celebrity. The comparison shows that *Shakira*'s page had a 47% growth in N_f, meanwhile, this growth value for *Adele* is almost half (28%). It can be concluded that however the number of fans of celebrities keeps to grow even during the period of inactivity, but due to the discussed point, there is a potential of a higher growth in the case of being active and publishing posts.

Our another observation shows that most of the inactive fan-loser users are personal pages which in contrast to the celebrities' pages, they lose fans in the period of being inactive. we take two pages *Jane Austen* and *William Shakespeare*, who are the two famous authors in ancient British, for example. *Jane Austen*'s page has about 3.1% of growth in phase2 with average 9 published posts per month, while *William Shakespeare*'s page has 2.5% of losing rate. In this case, even though we do not who is responsible for publishing posts for those pages, the active pages still show high possibility to attract fans compared with those inactive ones.

3.12.2 Impact of External Events

External events associated with professional users are influential factors on the visibility of users which provide as well a possibility to attract new fans in OSNs. Several studies focused on some major events that have clear impact on professional users such as political campaigns (e.g. elections) for politicians [71], sport events (e.g. Olympics, World-Cup, NBA and so on) for athletes and sport teams [72] etc. Nowadays, OSNs have become a major way to propagate information, such as governance brief announcement, personal perspective from celebrities, new products launched and so on. Every behavior and announcement can have a huge effect about the reputation of professional users, and the popularity variation on OSNs can also tell from this phenomenon.

Apart from the above-mentioned major events, there are also other specific events which impact particular professional user, such as a concert or a new album for a singer or a newly released movie for an actor/actress. Since each event usually is related to a specific user then the effect of events mapping on OSNs should be studied separately for each user.

To elaborate this impact, we select two sample users (*Aamir Khan* and *Allen Iverson* mentioned in Table 3.6) which have high growth rate without activity, and study the effect of external events on their popularity growth. *Aamir Khan* is a Bollywood actor and *Allen Iverson* is a retired American professional basketball player. These two users have no activity over the 14 months of the study in first phase, but their popularity trends in this duration (shown in Figure 3.20) includes major peaks. To find the reasons behind the sudden growth of the popularity in some dates, we manually followed the news about them and found a set of dates when there were some events about these users and their names appeared in the headlines of news.

For the case of *Allen Iverson*, his retirement announcement from basketball by himself (on October 30, 2013), and then officially by his team (on November 2013), and finally holding a ceremony for his retirement (on March 2014), were three main external events

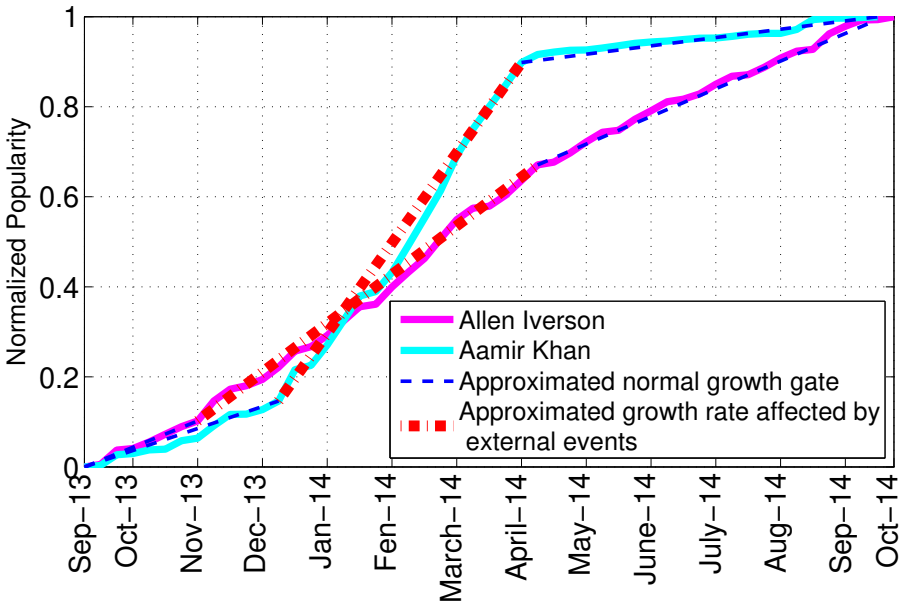


Figure 3.20 – Normalized popularity trends of Allen Iverson and Aamir Khan. The red and blue dashed lines show how the growth rate of their page's popularity has changed over time.

that occurred within 6 months and impacted his popularity trend shown in Figure 3.20.

Popularity of *Aamir Khan*'s page started to increase by a higher rate than before on December 2013 when a new Bollywood movie named *Dhoom 3* is released which he played the leading role on it. This movie got people's attention and broke many records in India and abroad. By following related news to *Aamir Khan*, releasing this movie was the main big relevant external event which attracted many people to follow his Facebook page even though there was no active post at that time.

To be more specific about the new release masterpiece which can also affect the popularity on OSNs, let us consider about the sample *Resident Evil* in Table 3.7. The series of *Resident Evil* movies have total 6 movies. The fifth movie *Resident Evil: Retribution* released on September 2012 was considered to be second to last by public audience in these series. A lot of negative reviews and low ranks in different website have impacted on its reputation of movie. We can also see in table 3.7 that *Resident Evil*'s page grow really slow by only 2.9% in the first period. However, The sixth movie *Resident Evil: The Final Chapter* came out on January 2017, which is exactly in our second phase. As the final released movie in the series of *Resident Evil* and plus, highly recommend reviews in different social media, *Resident Evil* got people's attention and grow about 258% rapidly during the second phase.

To summarize, external factors such as a big event, an announcement, being a hot topic in discussions or comments by other public people can significantly influence the popularity of a page. It represents normal users' like or unlike, which can easily be told by the growing

trend of *Fanpages*.

3.12.3 Other Influential Factors

In addition to the two mentioned factors, activity and external events, there are several other factors which can influence the popularity of a *FanPage* such as: (i) advertising campaigns both inside Facebook (using Facebook ads which recommend a page to people) as well as broadcasting the link of the *FanPage* on the official website of a brand or as a tag in posters, ads, etc. or encouraging people to follow a page, even with some prizes as enticement. (ii) multiple way of interaction with fans, such as live webcast, Youtube videos, replying fans' comments and reaction with their fans in real life and so on. (iii) smart activity such as publishing interactive posts which hugely engage people and encourage them to share a post on their wall page.

Studying the impact level of these factors is an interesting future direction of this study which will provide a better understanding of the phenomena of popularity in social media.

3.13 Conclusion

This section studied the evolution of users popularity in online social networks with a focus on professional users (e.g. companies, celebrities, brands, public figures, and etc.) and the variation trend on Facebook *Fanpages*. Toward that end, around 8K of the most popular professional users' *Fanpages* data have been collected over a period of 4 years. To be easy to identify, 28 chosen months have been studied mainly for population evolution part in this work. In addition, users' published posts also have been collected in the same time period, which eventually provided a popularity dataset including around 64 millions posts and 38 millions popularity snapshots.

The experiments conducted on this data reveal some interesting results. We observed that we can categorize users based on their popularity trends into two main groups of fan-losers and fan-attractors. In addition, we based on four different patterns in the first 14 months (phase1) of popularity evolution were identified. The popularity trend of same dataset evolves to five patterns in the last 14 months (phase2) during the study period. Moreover, we discussed the details about the *Fanpages* in different patterns, including categories, number of posts and activity level.

Next, we tried to understand the reason behind different popularity trends and why *Fanpages* changed their popularity patterns. We found several influential factors on the popularity trend of users, such as the level of activity (number of posts published, events associated with users, and interacting with followers), as well as external factors such as advertising campaigns, being temporary hot topic, etc. Simply being famous and celebrity (even a inactive one) can keep a user popular with a growth in the popularity value, but being an active user can substantially enhance popularity. The findings from this study provide a comprehensive view on professional users' popularity evolution, and reveal the influence of different factors on it, including external events and the activity level of users.

Several future directions can be taken to expand the result of this study. First of all, the two last sections have analyzed only professional users on Facebook. The popularity

analysis of these users on other major social network, (e.g. Twitter, Instagram, etc.) can be identified as more interestingly try in order to understand what is the difference of the popularity evolution for particular users across social networks. Secondly, this study mainly investigates the impact of activity and external events on the popularity evolution. There are several of reasons for *Fanpages* to evolve which could be very interesting to look at other influential factors, such as specific strategies, users' preference in social media, social engagement of users to the followers, etc. Lastly, the outcome of this type of researches can be taken as a guideline for professional users to enhance their success in social media. Preparing a comprehensive list of suggestions per user or sector can be considered.

Part II

Engagement Prediction on Social Media

Chapter 4

Who Will Like the Post? Predicting Likers on Flickr

Contents

4.1 Abstract

Reacting to a published post on a social media is one of the main activities of users which can happen in different forms comprising to like the post, leave a comment or reshare it. Finding a way to predict the size of user's future interactions and more interestingly identifying the users who are going to react to a post are the two important research topics which benefit different domains from efficient advertising campaign to enhanced content delivery systems. In this paper, we aim to predict the users who are going to react to a newly published post in future. Toward this aim, we implement a novel approach based on Point-wise Mutual Information (PMI) which derives users latent similarities from their interactions' log and exploits them to predict future interacting users. The proposed method is evaluated using a large dataset of Flickr including 2.3M users and 11.2M published photos. The empirical findings support the idea of employing interactions' log to detect future likers of posts by achieving noticeable prediction results for the tested dataset. Moreover, the analysis of the prediction task implies that likers prediction for the photos of publishers with a high number of followers and engagements is more accurate than the other publishers' photos.

4.2 Introduction

A great portion of the fast-growing research activities on social media has been devoted to the analysis of the data, which is available in these networks and more specifically the analysis of information propagation, users' characteristics, and engagements prediction [73] [74]. Users are the main actors of social networks who publish posts as well as reacting to the published posts by other users in various forms such as like, share or leaving comments. Users reacting to the posts on social media, are called *reactors* in this study. Reactors play a substantial role in information propagation and popularity of a post [8] [9].

The total number of engagement on a post shows the number of reactors, also known as the popularity number. Predicting this value and its involved reactors are two significant prediction tasks, which supply valuable information for many applications such as providing better solutions for content placement in networks, more efficient advertisement campaigns, and providing accurate recommendations. Among the existing efforts on these two prediction tasks the first one, predicting popularity size, has been inspected many times [10] [11]. However, identifying the users, reacting to the post has been neglected.

The key aim of this study is to identify future reactors of a post using of the prior information acquired from users' interaction log. Towards this aim, we have implemented a framework based on Point-wise Mutual Information (PMI) inspired by the $Word2vec$ language model [1]. Word2vec is a language model which derives word embeddings considering the co-occurrence of words in a *window* of vocabularies of size w. The proposed model in this study exploits different lists of users who have reacted to the published post via a *like* (marking the post as favorite) called *like sequences*, and computes the engagement probabilities of users on a newly published post. Since the reaction type in this study is specialized by *like*, we refer to reactors by *likers* term from now on. Using like sequences, we consider the co-occurrence of users in a window of size w to measure point-wise mutual

information between users. Considering users' co-occurrences in a window helps to discover the latent relation between them representing their similar preferences and favorite aspects which are not directly comprehensible from their friendships or profiles. In our method, PMI values show the strength of users' latent similarities in terms of their favorite contents.

We build a graph of users and their interactions, where nodes represent the users, and edges reflect the engagement probability of users on each of the other's posts. In order to build users graph, we consider three different approaches indicating three types of users graphs which differ in the type of links between users (directed or undirected) and in how the weight of these links is computed. The computed PMI value between two users is assigned to the weight of the edge between them. Given a new published post, we use the created graphs to find l users possessing the strongest links to the post's publisher representing the future likers of that post. These l users are the *l-nearest-neighbors* to the publisher, who are selected based on the PMI values between them and their neighbors, which are the most probable users who will like the post in future.

Besides the prediction of likers merely based on the publisher, we assume the availability of a prior-knowledge about k early likers of a post in addition to its publisher in order to take advantage of this knowledge and improve prediction results. In this case, we choose the l-nearest-neighbors from the neighbors of all k early likers. Prediction results are compared for different k numbers.

The main contributions of this study are:

i. We propose a novel approach to identify users who will react to the post by extracting users' latent similarity without using hand-crafted features.

ii. Although the likers of a post are not limited to its publisher's friends, comparing the prediction results when future likers are chosen from *all neighbors* versus from *only friends* shows that friends' interactions are more predictable than those of non-friends.

iii. We found that taking advantage of the *window* idea [1] to compute PMI values helps to predict more accurately.

iv. Our experiments reveal that future likers of a post are more dependent on the publisher of the post than early likers.

v. We identified number of followers and number of engagements of publishers as the most important properties that provide a better success rate in predicting future likers.

The rest of this chapter is organized as follows: The proposed methodology is presented in Section 5.2. The evaluation results of prediction and characterizing the successful publishers are discussed in Sections 4.4 and 4.5, respectively and Section 4.6 concludes the study and points avenues for future research.

4.3 Prediction Methodology

People interactions on the published posts produce temporal lists of users, representing the order of their interactions in different timestamps starting just after the published time.

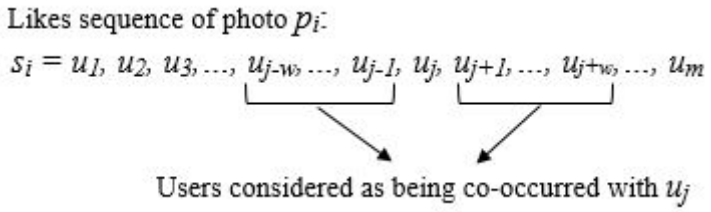


Figure 4.1 – Co-occurrences are computed for each user in the like sequences with her surrounded users, placed in a window of size w from two directions.

The main goal of this study is to design a model that is able to predict the potential users who will interact with a new published post[1] having prior-knowledge of the post's publisher and its early interactors. Taking advantage of PMI and inspired by Word2vec model [1], we implement a novel model to extract users latent similarities and their associations from their interaction logs.

4.3.1 Users Similarity

Given a social network with a set of N users (U), engagement of users on a given post p_i will be shown as $s_i = \{u_j \mid u_j \in U, j = 1, 2, ..., m\}$ and called interaction sequences (s_i), where m is the number of interactors on the post p_i, and index j refers to the index of users in a temporal order. Therefore, as shown in Figure 4.1 each s_i is a subset of users who reacted to the post p_i with the publisher of the post in the first place of the sequence (u_1).
In a dataset of P published posts, interaction sequences over those posts will be presented as the collection S where $S = \{s_i \mid i = 1, 2, ..., P\}$. As mentioned, each s_i includes the interactors of the corresponding post, p_i.

PMI values are computed for each pair of users as follows:

$$PMI(u_i, u_j) = \log \frac{P(u_i, u_j)}{P(u_i)P(u_j)} \tag{4.1}$$

Where $P(u_i, u_j)$ is the probability that two users u_i and u_j have co-occurred in interaction sequences. $P(u_i)$ and $P(u_j)$ are the probabilities that u_i and u_j appeared in an s, respectively. To compute $PMI(u_i, u_j)$, $P(u_i, u_j)$ is the first requirement which needs the number of users co-occurrences. In order to show the impact of *window* concept inspired by Word2vec on computing PMI values, we measure $P(u_i, u_j)$ in two approaches:

i. *Publisher-liker adjacency*: Co-occurrence is defined as the number of times that u_i is the publisher of a post and u_j is the user who reacts to that post.

ii. *Window adjacency*: This approach uses a window inspired by Word2vec, to compute the PMI values between user pairs. In this approach, we consider the co-occurrence of

[1] we call them likers through this study.

users in a window of size w, shown in Figure 4.1. This means that w users before and w users after u_i in the like sequence are considered as the users who are co-occurred with u_i.

Computed PMI values are assigned to the weight of the edges in the aforementioned user graphs. As already stated, the interaction graph of users is defined by considering users as its nodes. The edge between each pair of nodes is defined when one of the users reacts to the post published by the other. $PMI(u_i, u_j)$ is assigned to the weight of the edge between u_i and u_j in the interaction graph. The results are presented from the output of the following three approaches which are considered to build the activity graph:

- *Directed Publisher-Liker (DPL)*: The edges of this graph are directed and $PMI(u_i, u_j)$ is assigned to the edge which goes from node u_i to node u_j, if u_j reacts to u_i's post. PMI values are computed by aforementioned publisher-liker adjacency.
- *Undirected Publisher-Liker (UPL)*: The edges are undirected and the weight of the edge between u_i and u_j is the sum of the weights of the two directed edges between these two nodes in the previous approach (DPL).
- *Undirected Window (UW)*: This approach exploits *window adjacency* to compute the weights of the edges. Since the window adjacency considers different subsets of users from interaction sequences in which their relationship is not necessarily publisher-liker, the graph cannot be a directed one.

Two DPL and UPL approaches which use conventional definition of PMI are considered as the baseline methods to compare with UW approach where it uses new definition of PMI between two users under *window adjacency*.

4.3.2 Prediction Model

Here we describe in detail how our proposed method will identify the likers of a published post based on the users' latent similarities. Although PMI is widely used in prediction tasks on RS and NLP models, to the best of our knowledge, there is no previous study addressing the prediction of future engaging users using PMI and without hand-crafted features.

Due to the successful studies on predicting the popularity of posts by exploiting the information of early interactors [10] [39], we will also take into account the information of k early interactors of each post as a prior-knowledge and predict upcoming likers based on those earlier ones.

Given k early likers, we spot these k nodes on interaction graph, find the neighbors of each node, and make a collection of k nodes' neighbors. For each node in the collection, we compute the average weight of the edges between that node and k early likers. To identify future likers, we first sort the nodes available in the collection based on their already computed average weights and choose l top nodes with the highest weights referred by l *Nearest Neighbors* (l-NN). As the weights of edges are PMI values, the strongest edges imply the highest values on PMI. We select l-NN in two manners, choosing them from *all*

neighbors of k early likers like what described above as shown in Equation 4.2, and choosing them from *only the friends* of early likers according to Equation 4.3.

$$l\text{-}NN(k) = l\text{-}MAX(\frac{1}{k}\sum_{i=1}^{k} PMI(u_i,:)) \tag{4.2}$$

$$l\text{-}NN(k) = l\text{-}MAX(\frac{1}{k}\sum_{i=1}^{k}(PMI(u_i,:) * Friends(u_i,:))) \tag{4.3}$$

Where l is the number of chosen neighbors, k is the number of early likers as input, PMI is the matrix of PMI values, $PMI(u_i,:)$ is a row of PMI matrix indicating the PMI values between u_i and other users, and $Friends$ is the binary friendship matrix[2] and $*$ operation is the element-wise multiplication of two PMI and Friends matrices. The summation sign in both formulas applies an element-wise summation over the PMI matrix's rows belong to the k early likers. The average of this summation, which it is also element-wise average, is the input of $l\text{-}MAX$ function as a vector. This function selects l indices from the input vector having the highest average PMI values as l future likers. In Equation 4.2, future likers are chosen from all neighbors of early likers but in Equation 4.3, they are chosen only from the friends of early likers where it is achieved by multiplying PMI matrix by the binary friendship matrix. The two l-NN equations will be used to choose future likers based on early likers where the connection between users are defined according to the three approaches, DPL, UPL, and UW.

Next, we evaluate our proposed model by using a large Flickr dataset and present the outcomes of the prediction based on the different presented approaches.

4.4 Evaluation and Results

This section evaluates the proposed prediction method and presents the dataset information used in the evaluation as well as the results obtained from the experiments.

4.4.1 Dataset Description

To evaluate the proposed model of likers prediction, we used a Flickr dataset [75] including more than 11M photos and the activity history of 2.3M users for 100 days. User reactions to the photos in this dataset are indicated by marking them as favorites. In this study, we will refer this action by *like*, and the interacted users by *likers*. Table 4.1 shows the characteristics of the dataset and the values of its different attributes.

Since our method is based on the photos' like sequences, we consider those photos that have at least 30 likes to have enough length to apply the aforementioned idea of the *window*. Applying this filtering leaves the dataset to include 128k photos where each photo has the minimum number of 30 likes. In addition, to produce the reliable users' co-occurrence probabilities, a minimum frequency of likers is required. To fulfill this requirement, we pick only the users who have appeared at least 50 times in the dataset called *active users*.

[2] $Friends(u_i, u_j)$ is 1 if u_i follows u_j otherwise it is 0.

Table 4.1 – The Flickr Dataset Characteristic

Attribute	Value
#Photos	11.2M
#Users	2.3M
#Photos with ≥ 30 favorites	128K
Avg(#favorites) of ≥ 30 favorites	61
Median(#favorite) of ≥ 30 favorites	45

The thresholds for the number of likes and number of user's repetition are adapted from previous studies [60] [61]. The dataset is divided into two parts, namely train and test datasets, with 70% and 30% volume of the dataset, respectively. The train dataset is used to compute the PMI matrix and test dataset is exploited to predict the future likers.

4.4.2 Future Likers Prediction

The principal goal of the present study is to predict the group of users that have more probability to react in near future to a given post, where the prediction model will use only users' engagement log. To this end, we first have to choose the window size in the prediction model. We examined different values of w, but due to the space limit, we present the result for our model with $w = 10$.

PMI between users is computed from like sequences available in the train dataset, through the neighborhood of size w using the Equation 4.1. To predict future likers, the number of early likers (k) is set to vary from 1 to 20 in the two previously described Equations 4.2 and 4.3. By assuming to be aware of k early likers, we find l top users who are most expected to like a given post as the future likers of that post. Selected l users have the maximum amounts of average PMI values with early likers, representing the closest and similar users to the early likers. In order to set the value of l, we need to know the potential number of likers that will be predicted for each photo. This number comes from the number of active users in each like sequence. Because non-active users are already eliminated from the like sequences due to their repetition less than 50 times in the dataset.

Considering that this number of active users is different for each like sequence, we fixed the maximum number of likers to predict (l) to 20, which is the average number of active users in the like sequences of the train dataset. The prediction phase is conducted over the test dataset. We applied l-NN function using three approaches mentioned in section 4.3.1 and chose future likers. Prediction result is represented in two aspects, photos precision and likers precision.

Photo Precision

Photos precision indicates the portion of photos which at least one of their future likers out of 20 ($l = 20$) has been predicted correctly (called *predicted-photos*). Figure 4.2 shows the photo precision of different approaches along the Y-axis for the different number of early likers (k) along the X-axis as prior-knowledge. As it shows, we performed prediction of likers

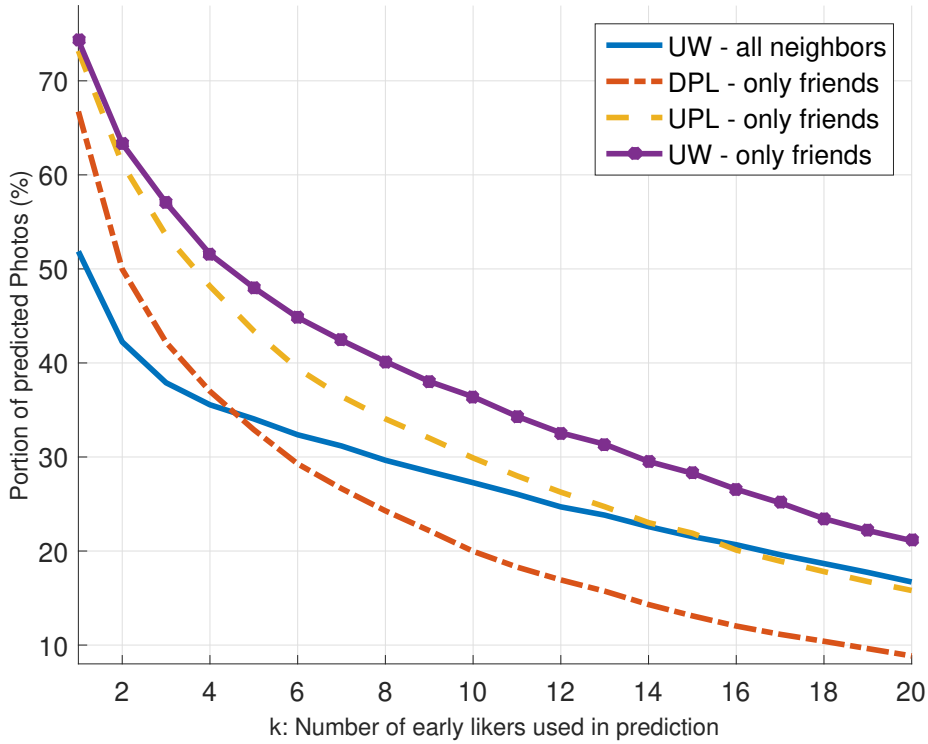


Figure 4.2 – The portion of photos with at least one correct prediction in their future likers for different k numbers of early likers. Choosing likers from friendships improves the prediction results.

using *DPL*, *UPL*, and *UW* in two categories, first choosing likers from all neighbors, and second from only friends. Since choosing likers from all neighbors shows significantly lower accuracy than choosing them from only friends, we circumvent to present the results of it in all three approaches. However, we present the result of *UW* from this category as the best representation of this group only to display its low accuracy. It is somehow expected that looking for future likers among friends provides more accurate predictions. First, because most of the users on any social network mark a post as liked due to their friendship with the publisher of that post, without considering the content of the post. Second, choosing future interactors from the entire users without restricting the search space, especially in such big datasets will not be intelligent and applicable. We also examined the random selection of likers as a baseline method to compare with our proposed approaches. But due to its very inaccurate result, we avoid presenting it.

As Figure 4.2 depicts, among four examined approaches, *UW* and UPL when they choose likers from only friends, can predict likers for the higher number of photos. In addition, it shows that by increasing the prior-knowledge about early likers (k along the X-axis) and subsequently choosing future likers based on them, the portion of predicted photos has been substantially declined. The highest number of correctly predicted photos is when the value of k equals 1, which is the case when we choose the nearest neighbors

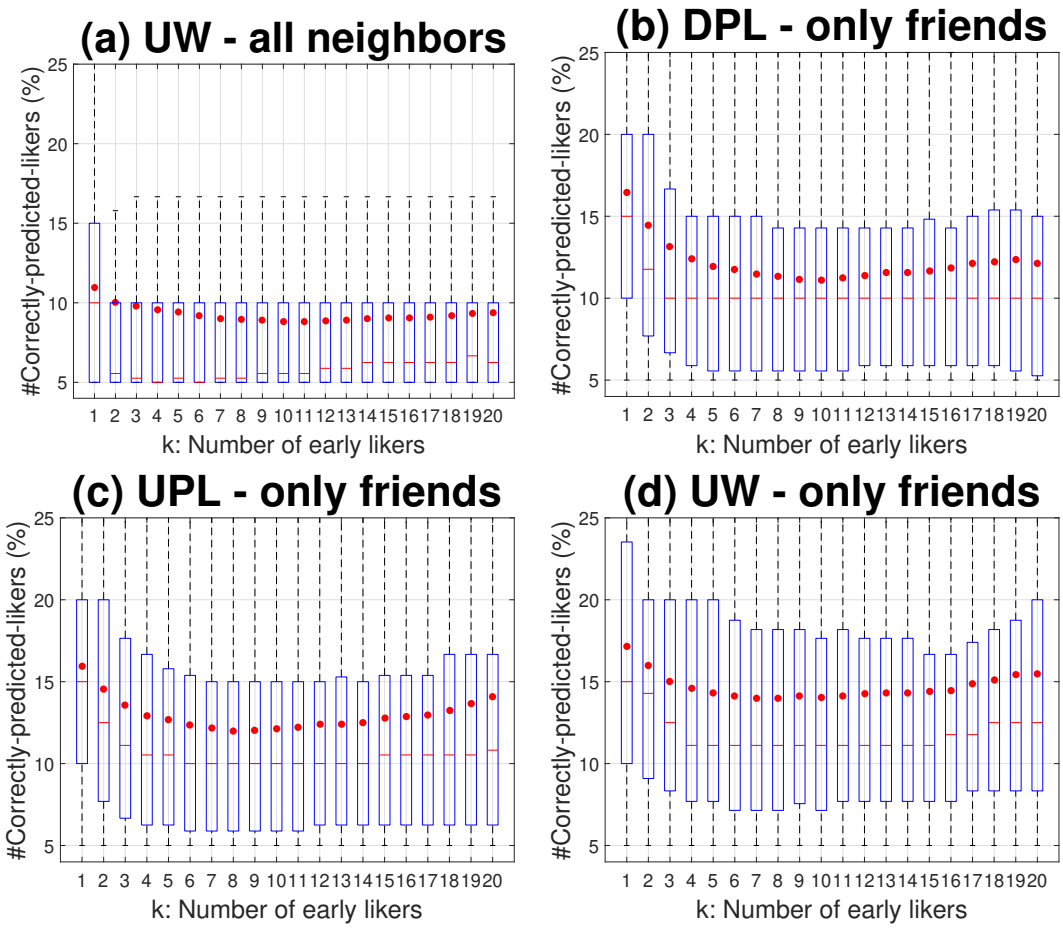


Figure 4.3 – Distribution of #correctly-predicted-likers of photos with different values of early likers (k). (Red dots show the mean values).

to the first user in the like sequence of a post who is the publisher of that post. It can be interpreted that the future likers are practically dependent upon the publisher. In other words, being aware of more early likers than publisher not only will not enhance the prediction precision but also will inject noisy data which leads to an inaccurate selection of likers.

Likers Precision

Likers precision is defined for each photo separately and indicates the portion of l predicted likers that are predicted correctly. Figure 4.2 shows only the quantity of photos which at least one of their future likers is predicted correctly using different approaches, without representing the quality of prediction. To identify the quality of prediction which indicates likers precision, we inspect precisely the number of likers which are predicted correctly for each photo (#correctly-predicted-likers). Figure 4.3 presents the distribution of these

numbers along the Y-axis (in percentage) for different k values. According to this plot, *UW* by choosing likers from all neighbors has the lowest mean value (presented by red points) of #correctly-predicted-likers where it is almost around 10%. On the other hand, *UW* with choosing likers from friends shows the best results such that first, the mean of #correctly-predicted-likers remains almost around 15% for different k numbers (against to *UPL* and *DPL* which drops) with the highest value at $k = 1$ and second, the distributions in each value of k show the higher numbers of predicted likers in *UW - only friends* than other approaches. *UPL* and *DPL* have almost similar distributions of predicted likers as well as similar mean values in different numbers of k.

As we observed in both Figures 4.2 and 4.3, the number of predicted photos (photo precision) and number of correctly predicted likers (likers precision) have their highest values in $k = 1$. It practically signifies that unlike the popularity size prediction problem [11], the prediction of future likers depends on the publisher of a post more than other early likers. Due to this point, in the following section, we will focus on the results of $k = 1$ where the photo precisions in Figure 4.2 are 51%, 66%, 73%, and 74% for *UW - all neighbors*, *DPL*, *UPL*, and *UW - only friends*, respectively. The purpose of this focus is to choose the best prediction approach among the presented ones in the elaborated presentation of #correctly-predicted-likers distributed in Figure 4.3 when $k = 1$.

4.4.3 Publishers as Predictors

As mentioned earlier, this section concentrates on presenting the results of the prediction on $k = 1$ which leaves the prediction problem to find future likers based on only publisher. To elaborate the results obtained from different approaches, we compute the precision of prediction for each photo (p) called $Precision_p$ as follows:

$$Precision_p = \frac{\#correctly - predicted - likers_p}{\#likers - to - predict_p} \tag{4.4}$$

Where #correctly-predicted-likers$_p$ is the number of likers of the photo p who are predicted correctly, and #likers-to-predict$_p$ is the number of photo p's likers. To provide simpler representation, we group $Precision_p$ values into ranges. Figure 4.4 displays the distribution of photos' precisions ($Precision_p$) computed from the results of four prediction approaches. In this figure, the first bar in the range of 5-10%, which belongs to *UW - all neighbors* approach, shows that this approach can predict only 5 to 10 percent of likers correctly for 23% out of 51% *predicted-photos*, 10 to 15 percent correct prediction for 18% and so on.

Comparing different approaches reveals that *UW - all neighbors* has the majority of its correctly predicted photos in the range of 5-10%. It means that only 5 to 10 percent of likers are predictable for almost half of the predicted-photos (23% out of 51%) using *UW - all neighbors* approach. Therefore, this approach not only has the lowest percentage of predicted-photos but also is not able to predict more than a few percentages of likers. Contrary to *UW - all neighbors*, the other three approaches perform better and the likers of the majority of photos are predicted by 10-15% and 15-20% precision using those three approaches. It implies that when likers selection is restricted to choose them only from

friends instead of from all neighbors, the precision of the results is substantially enhanced. The reasons behind this phenomenon are previously discussed in Section 4.4.2 as well.

From the three better performing approaches, *UW - only friends* outperforms *DPL - only friends* and *UPL - only friends* by resulting a higher number of photos with high $precision_p$ in likers prediction. As Figure 4.4 shows in the precision ranges higher than 20%, the number of predicted-photos by *UW - only friends* beats the others. Accordingly, it substantiates the success of this method in predicting high number of likers. In summary, we found that prediction of future likers considering their relation with publisher provides a better result than with other k early likers. Restricting the prediction to choose likers only from friends instead of selecting them from all neighbors elevates the quantity of number of predicted-photos by more than 20% (from 51% in *UW - all neighbors* to 74% in *UW - only friends*), and the precision of likers prediction from low to high ranges.

Finally, *UW - only friends* succeeds to predict the higher amount of photos with higher precision of likers in compare to the other three approaches. As stated previously, this method exploits the co-occurrence of users in a window of size w, which makes it able to derive the latent similarity between users even if they have not interacted directly on the posts of each other. Consequently, considering a window to compute the co-occurrences of users helps *UW - only friends* to improve the precision of likers in the prediction task.

4.5 Publishers Analysis

As we observed in section 4.4.2 likers precision is different for each photos. In order to identify why some photos have more correctly predicted likers than others, we study the properties of their publishers. Looking at the prediction result shows that the *UW - only friends* approach produces the best outcomes (although *UPL - only friends* was very close). Thus we study the result of this approach to discover the common features of those publishers that likers of their photos are predicted more accurately. To this purpose, we investigate publishers properties studying their relationships, activities, and engagements. Four metrics are considered for each publisher: #followers, #followings, #activities (published photos) and #engagements (number of times that a publisher reacted to the photos of other publishers).

Earlier we defined two $\#likers\text{-}to\text{-}predict_p$ and $\#correctly\text{-}predicted\text{-}likers_p$ metrics for individual photos which are the number of users in p's like sequence and the true predicted likers of p, respectively. Now we will define the same metrics for publishers. Since each publisher has different number of photos in the dataset, we compute the average of these values for the photos of each publisher and associate them to the corresponding publisher as the average values of potential #likers-to-predict and #correctly-predicted-likers of that publisher. On the other side, to show how many photos of each publisher have at least one correct prediction of their likers, $Predict-frac_{pl}$ represents the percentage of the following fraction:

$$Predict - frac_{pl} = \frac{\#predicted - photos_{pl}}{\#published - photos_{pl}} \tag{4.5}$$

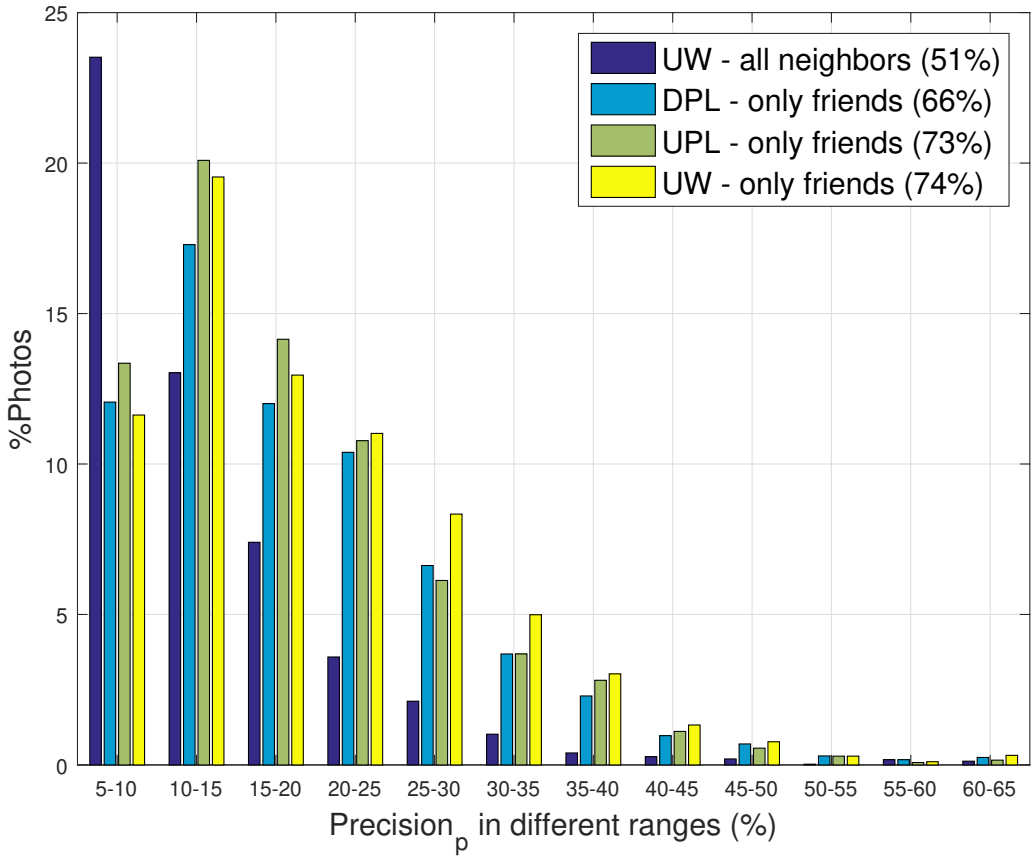


Figure 4.4 – Distribution of predicted photos (Y-axis) over the different ranges of $precision_p$ which is computed per photo separately in $k = 1$. (the portion of each bar is from the percentages shown in the legend)

Where #published-photos$_{pl}$ is the number of photos published by the publisher (pl) and #predicted-photos$_{pl}$ is the number of her photos with at least one correct predicted liker.

Figure 4.5 compares publishers in terms of their $predict - frac_{pl}$ shown by color, the number of predicted photos shown by the size of the circles, the average number of predicted likers in the Y-axis, and the average number of likers to predict along the X-axis. From the perspective of the quantity of predicted photos, the most successful predictions belong to the publishers with the higher values of $predict - frac_{pl}$ represented by blue (and darker) colors and the higher #predicted-photo presented by larger circles in Figure 4.5. In addition, from the perspective of prediction quality, the most successful predictions are associated with the publishers whose average number of correctly-predicted-likers are high. We call them the high-predictable publishers, located on top of the plot along the Y-axis. We used a heuristic to find a reasonable number of high-predictable publishers. We intuitively filtered publishers by selecting those which own #predicted-photos of more than 20, or have $predict - frac_{pl}$ value greater than 80% or those whose the average number of correctly-

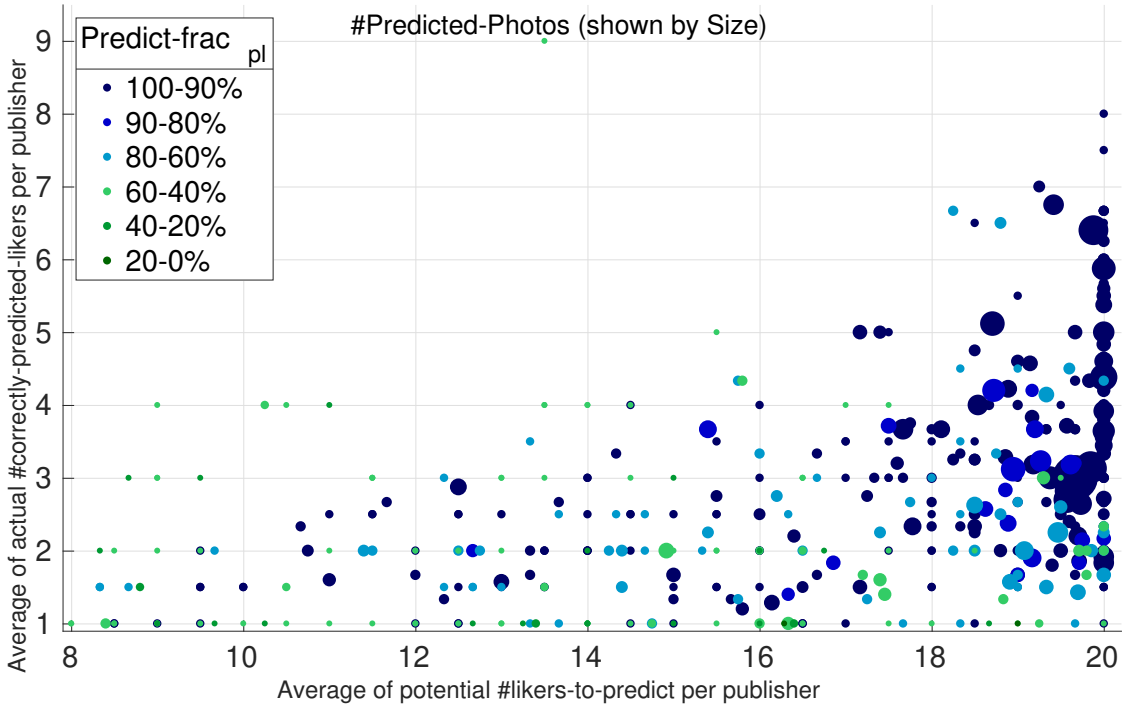


Figure 4.5 – Four-dimensional representation of publishers consisting avg. potential #Likers-to-Predict (X-axis), Avg. #correctly-predicted-likers (Y-axis), actual #predicted-photos (shown by the size of circles), and predict-frac (shown by color). The bigger size of circles shows the higher number of predicated photos and vice versa.

predicted-likers is more than 5. Among the selected publishers, the ones who met three applied filtering conditions are grouped as the highly predictable publishers with 22% of selected population, and the others who met only two or one of the filtering conditions are grouped as lowly predictable publishers with 78% of the whole filtered publishers.

To determine the characteristics of these two groups and to identify the influential factors in the success of highly predictable publishers, we compare the four previously mentioned metrics of the publishers in those two groups in Figure 4.6. The value of each metric is the average value in this diagram. It shows that high-predictable publishers have significantly higher values for their number of followers and engagements than the low-predictable ones. These values are almost twice bigger for high-predictable publishers. #followings of high-predictable publishers is almost 50% greater than the value of the same metric for the low-predictable publishers. However, the average amounts of the published photos (activities) by those two groups are almost equal, which indicates that a user's activity-amount regarding publishing posts is not a referable metric to determine the predictability of her photos' likers.

These observations reveal the substantial effect of a publishers' high number of followers and high number of engagements on achieving a successful prediction of her content's likers,

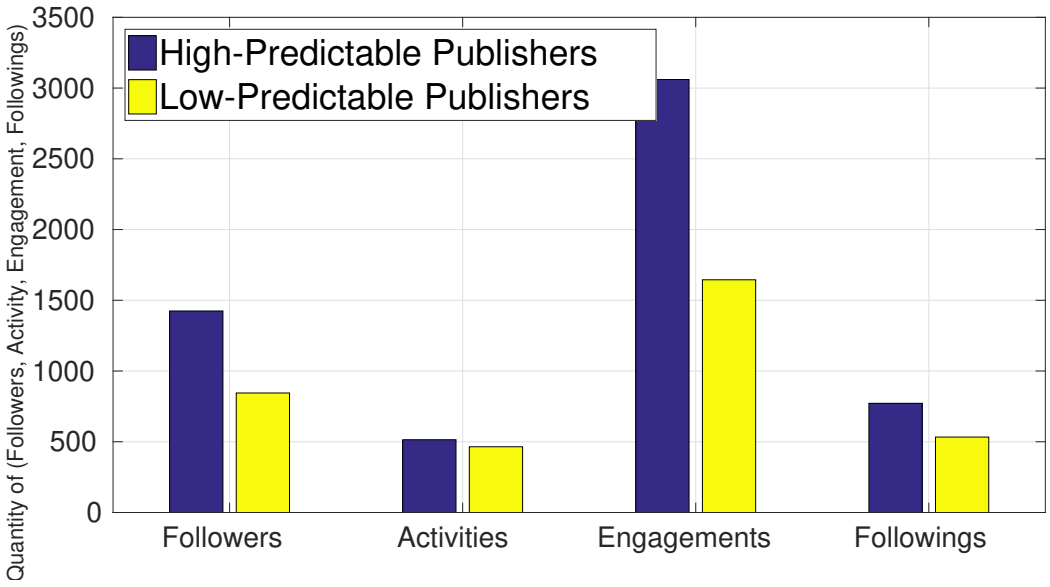


Figure 4.6 – Comparison of high-predictable and low-predictable publishers in terms of their average values of #followers, #engagements, #activities (published-photos), and #followings.

employing only user's interaction history. This means that the future reactions to a post published by a publisher with high #engagements and high #followers are more predictable. In addition, since we predict likers by exploiting the PMI closeness in *UW-only friends*, we can infer that using the latent similarity of users derived from their engagement history can produce more reliable results for those users with a high number of followers and engagements to use in this prediction task.

High correctness of prediction for the posts of publishers with high #engagements implies that engaging a user in the posts published by other users helps to reveal their common preferences as well as helps to make other users more predictable in reacting to her future posts. On the other side, publishers with more followers increase the probability of accurate prediction results because their contents will probably get a high number of likes. The results of this section illustrate that exploiting activity sequences can effectively extract trustworthy latent similarities between active users to employ in predicting future likers especially for the publishers with high number of followers and engagements.

4.6 Conclusion

This study sheds light on the interesting topic of predicting the users who are most likely to react to the posts published in social media. A novel model based on PMI and inspired by Word2vec was implemented to extract users' latent similarity. The similarity of users is exploited to predict the future likers of a post based on the information of the post's

publisher as well as its early likers.

Our findings disclose that considering users adjacency under a window of neighborhood reveals users hidden similarities and leads to more precise PMI values. As well as, we found that predicting future likers of a post is considerably correlated to the publisher of that post than other early likers. We studied in details the output of prediction model from photos precision and likers precision perspectives. Evaluation of experiments over a large Flickr dataset confirmed the ability of the proposed method to identify future likers of Flickr's posts, especially those published by super interactive publishers.

Although the study has reached the worthy results, it is limited by the lack of homogeneous data from other social networks to generalize the results for larger number of social network platforms. The proposed prediction approach can help advertising campaigns, recommender systems, and content placement controllers by providing prior-knowledge of future engaging users. Further research could include improving the outcomes of the proposed method by augmenting users' information to the content of posts as well as fine-tuning this technique to extract users' latent relations and preferences. Some of these goals could be realized by applying this method to distinct datasets.

Chapter 5

User Reactions Prediction Using Embedding Features

Contents

5.1 Abstract

By the massive available people data in social media, many digital service providers exploit widely this information to improve their services by predicting future requirements of their customers. This prediction mainly needs to study users' previous behavior and interactions and identify their preferences to provide rigorous recommendations that fulfill their requirements more favorably. Meanwhile, experiments show the prediction methods which exploit representation learning instead of traditional hand-crafted features accomplish better results and more precise predictions.

In this study, we take advantage of representation learning method to predict user's future interactions by extracting users embeddings from their reactions history and exploit them in predicting future reactions. In this approach, users embeddings are used in a neural network designed with one-hidden layer and a softmax function in the end layer in order to predict users reactions. The proposed method is evaluated when user embeddings come from two different sources; users reactions history and random walks on the user network. The performance of the method has been evaluated by using a large Flickr dataset including more than 2M users and 11M users reactions sequences. The results show outperforming the prediction method when it uses the history of user reactions to derive user embeddings.

5.2 Introduction

Nowadays, every aspect of human life has been widely affected by social media. An enormous amount of data is uploaded to Online Social Networks (OSNs) every day which is analyzed and employed to improve the user services provided by those networks. Among the different research directions through analyzing social media data, predicting user's future behavior in order to serve efficient user services is one of the most attractive studies. User behavior prediction plays a key role in a wide range of applications such as recommender systems, content delivery networks, advertising campaign, election results prediction and the list goes on. User behavior comprises her preferences, her interaction [1] type such as post, comment, share, like, and so on. In this study we will focus on predicting user behavior in terms of her interaction on social media.

Once a post is published on a social network, depending to its interestingness for other users, it could attract a particular amount of user interactions. Predicting the amount of user interactions and more interestingly the users who will react to that particular post are two main trending research tracks. Some studies have focused on predicting the final size of the popularity of a content to provide a vision of trending content [40] [10]. But some others have targeted more details and tried to predict the users who will make a content popular in addition to its final popularity size [41] [42]. This study will focus on predicting the users who will interact with a newly published content in near future.

Usually prediction tasks on social networks are based on learning methods which need features to be used in the model. Finding the most efficient features that provide more accurate prediction is always one of the main challenges on using conventional learning

[1] In this study, the word interaction is used interchangeably with reaction.

methods such as classification, clustering, regression, etc [11] [10]. Recent success stories of deep learning in extracting embedding features have led to exploit this method in different data mining tasks by skipping the manually feature extraction phase [45] [76]. Aligned to this research direction, this study aims to take advantage of representation learning in order to learn user features without require to hand-crafted features. It will use users interaction history to extract their embedding features and exploit those features in prediction whether they will react to a post published by one of them.

As users react to posts that they are more interested or have friendship with the posts' publishers, a sequence of users interactions and their co-occurrence in that sequence can represent their common preferences and interests. The proposed model exploits users' reaction log as the input of the Word2vec model to derive user embedding features. Depending on the type of social network, the reaction of users can be in the form of re-share, like, or comment on the post. Since we use Flikr data in this study, *marking a photo as favorite* is considered as user reaction. Previously, some models such as Node2vec also have extracted user embeddings to exploit in prediction tasks such as multi-label classification [60] [47]. However, the input of Node2vec is random walks over the users network graph which are not applicable for our following purpose. Since our goal is to find such features that can represent users tendency to react to a post, their neighborhood in the graph emerging in random walks can not provide this tendency. Therefore, the hypothesis of deriving features from feeding interaction logs to the Word2vec model will be followed in this study which provides better features to take their benefit in interaction prediction task. Users embeddings are exploited in a one-hidden layer neural network with a softmax function in the end layer to predict users reactions. We compared the results when users embeddings come from the Node2vec model and from the reaction sequences.

Using users' reactions log to learn their embeddings and predict their future interactions are the main contributions of this study. Besides, the proposed model in this paper is general and can be applied to any social networks data. Our experiments show more accurate results compared to other existing approaches which can potentially be used in different recommendation scenarios.

5.3 Methodology

Given a post and its publisher, the aim of the present study is to predict users' reactions to that post. Depending on the type of social network, the reaction of users can be in the form of re-share, like, or comment on the post. We consider the prediction task as a probability function to decide whether a user will react to the post of a given publisher. To this end, we first extract user embedding features and then exploit them as input in a simple neural network with a softmax function in the last layer.

5.3.1 Reactions Sequences

Following the main objective of the study on predicting future reactions of users, we exploit users' previous reactions log to extract their embedding features at the first step of our model. In a social network with a set of N users (U), reaction of users to a given post

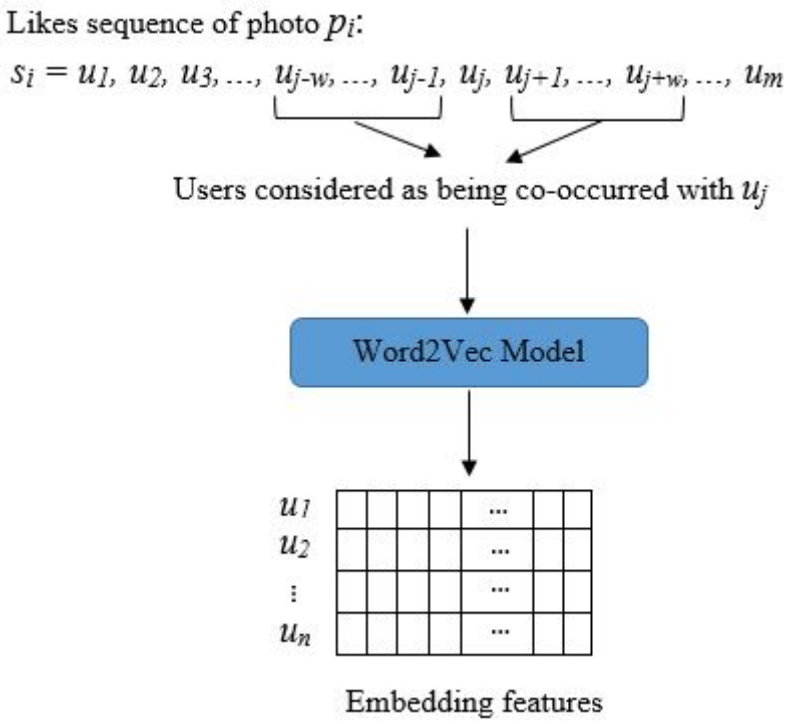


Figure 5.1 – Co-occurrences are computed for each user in the reaction sequences with her surrounded users, placed in a window of size w from two directions, are fed to the Word2vec model. the Word2vec Model supplies user embeddings.

p_i, published by u_i, will be shown as $s_i = \{u_j \mid u_j \in U, j = 1, 2, ..., m\}$ and called an interaction sequence (s_i). Where m is the number of interactors of the post p_i, and index j refers to the index of interacting users in a temporal order, shown in Figure 5.1. We put the publisher of the post in the first place of the sequence.

In a dataset of P published posts, interaction sequences over those posts will be presented as the set S where $S = \{s_i \mid i = 1, 2, ..., P\}$. As mentioned, each s_i includes the interactors of the corresponding post, p_i.

There are mainly two reasons behind the reaction of a user to a post. First, the relation of the user with the publisher of the post such as friendship and followership. Second, the user's interest to the content of a post which induces her reaction to that post. As our aim is to investigate the competence of users pair-wise relations in predicting their future reactions, we are supposed to achieve the second concern by extracting the likelihood of users' interests from their common reactions to the posts. Considering the mentioned points, we exploit users interactions sequences to derive user embeddings. In users' reaction history, the users' neighborhood illustrate their latent common tendency to react to the posts. Furthermore, since the neighborhood of users in a reaction stream can be a representation of cascading paths, user embeddings extracted from reaction streams will implicitly include cascading pattern between users as well. We use the Word2vec model

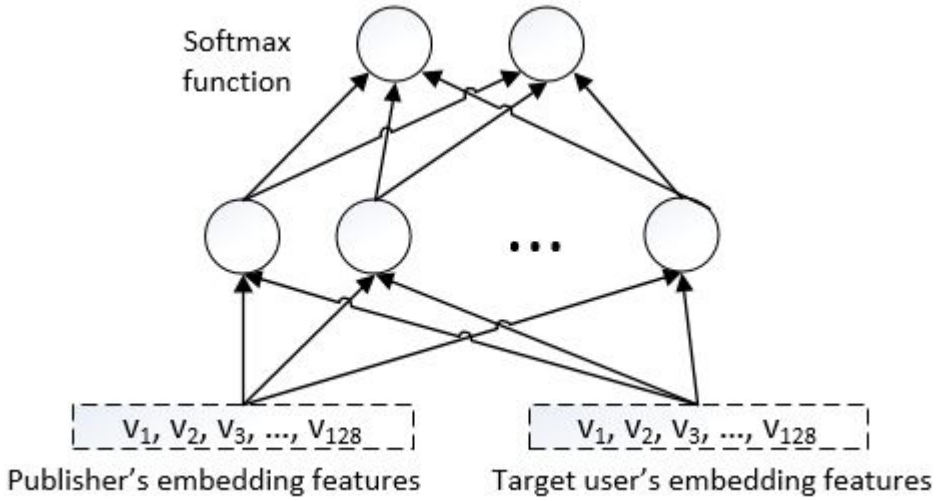


Figure 5.2 – One-hidden layer neural network with softmax function in the final layer to predict the target user's reaction probability.

in order to extract user embeddings. As Figure 5.1 shows, to generate user pairs required in the Word2vec model, we consider w users before and after each user in a reaction sequence to be paired with that user. Produced user pairs demonstrate the number of times that each two users are co-occurrenced in reaction streams within the window of size w. The high number of co-occurrences of user pairs indicates the more similar interests of them. Users pairs are fed to the Word2vec model and the model derives user embedding features as explained in subsection 2.4.2

5.3.2 Future Reactions Prediction

We aim to use user embeddings extracted from reaction sequences to predict who will react to a given post. To reach this goal, we have designed a simple neural network with a softmax function in the final layer, as shown in Figure 5.2. Given a post's publisher, the network will make decision about user's reaction to that post. Inputs of the network are embedding features of the publisher and a target user, whose reaction probability is going to be predicted. The embeddings are extracted from the Word2vec model as described in previous section. The middle (hidden) layer performs a dot product with two input vectors and their weights, adds biases and applies the Rectified Linear Unit (ReLU) activation function. Output of the softmax function in the last layer will be a two-dimensional binary vector. It represents the probability of the target user's reaction over the post of the given publisher. In our model, we will consider only users features to predict their future interactions without considering the content of the post.

As there is no specific method to determine the best number of layers and nodes for a neural network, we have tried different number of hidden layers as well as different number of nodes in the middle layer of the network and chose the numbers that provide

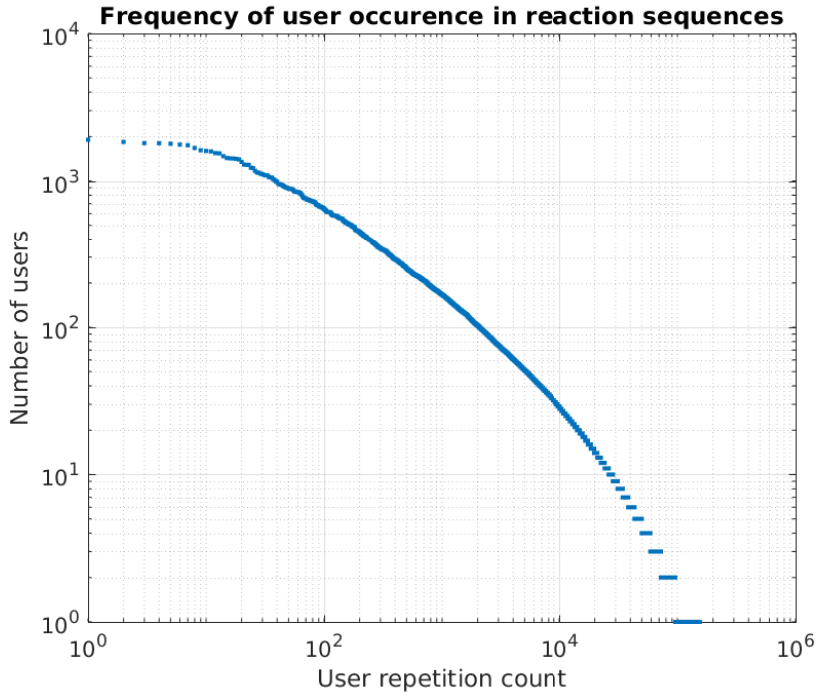


Figure 5.3 – The power-law distribution of users in reaction sequences.

outperforming results. The network is trained for different number of epochs of the data. Network uses Adam optimization method [77] implemented in Tensorflow[2] to update the weights.

5.4 Evaluation

This section discusses the dataset which is used for evaluation as well as the conducted experiments and the obtained results of the proposed prediction approach compared to a baseline methods.

5.4.1 Dataset Description

The Flickr dataset from [75] is used to evaluate the proposed approach of reactions prediction. The dataset includes more than 11M photos and 2.3M users' activity log for 100 days. As mentioned previously, the reaction of users can differ from a social network to another. In this dataset, user's reaction to the photos is referred by marking the photos as user's favorite, which we refer this reaction by *like*, and the interacted users by *likers* in this study. The dataset includes the followership information between users as well. The main characteristics of the dataset are shown in Table 5.1.

[2]https://www.tensorflow.org/

Table 5.1 – The Flickr Dataset Characteristic

Attribute	Value
#Photos	11.2M
#Users	2.3M
#Photos with $\geq$ 30 likes	128K
Avg(#likes) of $\geq$ 30 likes	61
Median(#likes) of $\geq$ 30 likes	45

We check the distribution of users repetition in reaction sequences and observe that it follows a power-law distribution shown in Figure 5.3, which is very similar to the power-law distribution of words in natural language. As our technique is very similar to the Word2vec model, and the Word2vec model is very successful to capture word embeddings, this similarity verifies our method to be suitable to capture user embeddings. Our main contribution is the idea of using natural language processing techniques to model users reaction behaviour on social networks.

Shown in Section 5.3.1, reactions to a post provide us a list of users who have liked that post. Since our aim is to learn users representations from the like sequences, we need to choose those photos from the dataset that have enough like sequence length to support the window concept in the Word2vec model. Therefore in our experiments we only consider photos which have at least 30 likes. This threshold number can be different depending on the datatset size. The final dataset includes 128k photos whose number of likes are more than 30. The dataset is divided into two parts. The first part, with 50% of the data, is used to learn embeddings. And the remaining part of the data is used to train and test of the designed neural network.

5.4.2 Likers Prediction Experiments

Model Configuration

As mentioned, 50% of the prepared dataset is considered as the input of the Word2vec model to extract users embeddings. We set the window size to 10 and the number of features to 128. Derived embeddings are used to predict the like prediction. As shown in Figure 5.2, we designed a one-hidden layer neural network with embeddings of publisher and the target user, whose reaction is going to be predicted, in the input layer, 8 nodes in the middle layer and two nodes in the final layer indicating the binary like probability. As there is no specific way to determine the number of layers and nodes in a neural network, we evaluated the results when we have more or less layers and nodes in hidden layer, however the one-hidden layer network with 8 nodes in the middle layer outperforms other configurations.

We consider the remaining 50% of the dataset to train and test the designed prediction network. Since the dataset is composed of photo's like sequences, we have only the users who have liked the photos. We needed to generate negative samples who did not like the given photo. To have a fair dataset, the number of negative samples for each photo

Table 5.2 – Hand-crafted user features used by SVM classifier to predict user's future reactions.

Feature	Description
$\#Likes_p$	Number of photos has been liked by publisher.
$\#Likes_u$	Number of photos has been liked by target user.
$\#Photos_p$	Number of photos that publisher has published.
$\#Photos_u$	Number of photos that target user has published.
#Reciprocal_likes(p→u)	Number of the given publisher's photos has been liked by target user.
#Reciprocal_likes(u→p)	Number of target user's photos has been liked by the given publisher.
#Mutual_likes	Number of photos from other users that both publisher and the target user have liked.

is considered to be equal to the number of users (likers) in that photo's like sequence. Negative samples for each photo are chosen from the publisher's friends and non-friends in the same portions that they are distributed in the like sequence.

Baseline Methods

We compared the prediction results with two base-line methods. The first baseline method is a Support Vector Machine (SVM) classifier [78] which uses some hand-crafted features. We choose this conventional learning method to compare with our method in order to provide a comparison between hand-crafted features and embedding features. Following the aim of this study which is prediction of user's future reactions using only previous reactions' history, we extract the user features listed in Table 5.2 to be used by SVM.

We use our proposed prediction network when user embeddings come from a different source, as the second baseline. In this method, user embeddings are derived from random walks over the user graph (Node2vec) [47]. We aim to compare the efficiency of user embeddings when they come from two different sources. This comparison will reveal whether our idea of extracting user embeddings from user reactions can benefit the prediction task followed by this study. We chose the same feature size and window length for both like sequences and random walks. Performance of the experiments is evaluated in different epoch numbers, learning rates, and batch size, and eventually we chose the numbers that provide high performance.

Experimental Results

For our designed network, we examined different values of learning rate and batch size to find the best configuration of the network. In order to avoid presenting different numbers of each parameter and their different combinations, we show the results of the best performing examined values of the parameters. Table 5.3 represents the results for the SVM classifier,

Table 5.3 – The performance of the proposed reaction prediction model in two different input sources.

Input of the model	**Precision**	**Recall**	**F1**
SVM classifier	0.501	0.512	0.506
Random walks	0.584	0.684	0.630
Reaction sequences	0.609	0.756	0.673

and our proposed approach when it uses like sequences embeddings and random walks embeddings. Three following metrics are considered to compare, *precision*, *recall*, and *F1_score*:

$$Precision = \frac{tp}{tp + fp} \tag{5.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{5.2}$$

$$F1\_score = 2\frac{Precision * Recall}{Precision + Recall} \tag{5.3}$$

SVM classifier has the lowest values for all metrics. It shows that the SVM classifier using the defined features could not be discriminative in this prediction task. The results for our approach presented in Table 5.3 come from a configuration of 0.001 for learning rate and 512 for batch size. *Random walks* shows the performance of the model using user embeddings extracted from random walks over users graph, and *Reaction sequences* represents the performance of the same model when embeddings come from reaction sequences. As we can see, *reactions sequences* approach achieves higher performance than *Random walks*, and both of them behave better than SVM classifier. *F1_score* for the prediction model using reactions sequences embeddings is 67.3%, better than *F1_score* for random walks embeddings which is 63%. Precision and recall metrics get their highest values with 60.9% and 75.6% respectively, in *reactions sequences* approach which are more accurate than the result of *random walks* method.

This implies that users' neighborhood in reactions sequences can represent their likelihood better where the likelihood concerns their tendency to react to each other's posts. The most reasonable explanation for performing reaction sequences embeddings better than random walks embeddings is that users' neighborhood in reaction sequences involves implicitly users' preferences similarity in addition to their friendship. While in random walks graph, links are the only things keep them to be neighbor. It proves our hypothesis of using users' reaction logs to discover their latent similarity in terms of reacting to each others posts.

As a representative parameter assessment, Table 5.4 shows precision, recall, and F1_score of the model in different examined learning rates and batch sizes when the model uses reactions sequences embeddings as input. As we can observe in this table, although the performance of the model obtains very close values in some configurations but when the learning rate is set to 0.001 and batch size to 512 it achieves its highest value by 67.3% F1_score. According to our observations, when learning rate is 0.1 the model's performance

Table 5.4 – The performance of the model (with reaction sequences embeddings as input) with different learning rates and batch sizes.

Parameter		Precision	Recall	F1
Learning rate	0.1	0.577	0.746	0.651
	0.01	0.669	0.550	0.604
	0.001	0.609	**0.756**	**0.673**
Batch size	40	**0.694**	0.422	0.525
	512	0.609	**0.756**	**0.673**
	1024	0.577	0.687	0.627
	2048	0.628	0.676	0.651

is not stable. By decreasing it to 0.001, the network's weights get updated slightly and it helped the network to converge after almost 9 epochs. In our dataset, we reached the best performance when the batch size is set to 512.

5.5 Conclusion

This study aimed to predict users future reactions (e.g. likes, comments, shares) on ONSs using their reaction history on the published posts. Toward this goal, the proposed model first extracts users embeddings from their reactions log and use them to predict future engagement of users. Reaction history of users comes from their previous engagement to the content published on social media. We took advantage of user embeddings out of reactions sequences, where users likelihood represents their close relationship or preference similarity, to predict their reaction when a post gets published by a publisher. Users embeddings are exploited in a one-hidden layer neural network with a softmax function in the end layer. We compared the results when users embeddings come from the Node2vec model and from the reaction sequences. The experiments show higher precision when users embeddings are derived from reaction sequences. It means that reactions sequences present better likelihood of users than random walks through users graph, in terms of revealing their potential probability to react to a post. Although we mainly focused on Flickr dataset, the proposed model is a general approach that can be applied to different social networks. As a future direction of this research, we will take advantage of user graph and draw out the subgraph of each reactions sequence in order to find an approach that can extract users embeddings with no need to random walks over subgraphs.

Chapter 6

Conclusion

In this thesis, we proposed several novel popularity characterization and prediction methods for users and content on social media. We have firstly done a comprehensive overview on the state of the art related to popularity and reaction prediction on OSNs.

The evolution of user popularity is modeled through a clustering method. Various popularity trends and patterns including fan-losers and fan-attractors are identified. Identified clusters are investigated from different perspectives consisting number of fan, business category, and activity volume. In an extended study, the popularity patterns are investigated in a longer period of time to examine their variation. We compared the results from the first study to the second and discussed the evolution process. Moreover, we detected the impact of most influential factors such as external events on the popularity evolution of users.

We proposed two novel models to predict the future likers of the post published in social media. These two models are generic and can be easily applied to any social network. The input of the both models is users' reactions history. One of the models is based on the PMI values computed for each two users. In this model, PMI values are the output of a new proposed method which is inspired by word2vec.

The other model predicts future likers using a simple and shallow neural network without requiring hand-crafted features. This model first extracts user embeddings from reactions log and then feeds them to the mentioned neural network classifier to predict the likers. The proposed model not only has the novelty of deriving and exploiting user embeddings in likers prediction task, but also outperforms the other methods especially the conventional learning models.

Our perspectives:

i. With regard to the recent success of convolutional neural networks (CNN) on different areas such as image processing and natural language processing, our perspective is to design a CNN with the same input of our model to examine its performance in reactors prediction task.

ii. We believe that taking content embeddings into account in the reactors prediction task and design a new model which has an input mixed of user and content embeddings will help to improve the performance of the prediction model. In this case, the content can be text or image. The model will firstly extract text/ image embeddings from the content and user embeddings from the reactions log and then will feed them to a CNN or a simple neural network (similar to the proposed network in chapter five) to predict future reactors.

iii. As mentioned in the state of the art section, popular content is very important for service providers. Therefore, one of the future studies of this thesis can be designing a comprehensive predictive model possessing two prediction steps. The model first predicts the final popularity status of a given content (such as a binary classifier). This step predicts whether the given content will become popular in near future or not. Depends on the type of the content, different text, image, or video models can be used. The second prediction step will predict the future reactors only in case the first step predicts the content as a popular one.

Bibliography

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[2] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq, "Paying for likes?: Understanding facebook like fraud using honeypots," in *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, 2014, pp. 129–136.

[3] G. Stringhini, M. Egele, C. Kruegel, and G. Vigna, "Poultry markets: on the underground economy of twitter followers," in *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, 2012.

[4] F. P. Barclay, "Political opinion expressed in social media and election outcomes-us presidential elections2012," *Journal on Media and Communications (JMC)*, vol. 1, no. 2, 2014.

[5] F. Giglietto, "If likes were votes: An empirical study on the 2011 italian administrative elections." in *ICWSM*, 2012.

[6] R. Farahbakhsh, A. Cuevas, and N. Crespi, "Characterization of cross-posting activity for professional users across major osns," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM, 2015.

[7] K. Nelson-Field, E. Riebe, and B. Sharp, "What's not to like?" *Journal of Advertising Research*, vol. 52, no. 2, pp. 262–269, 2012.

[8] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," *arXiv preprint arXiv:1405.1790*, 2014.

[9] E. Dubois and D. Gaffney, "The multiple facets of influence: identifying political influentials and opinion leaders on twitter," *American Behavioral Scientist*, vol. 58, no. 10, pp. 1260–1277, 2014.

[10] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014.

[11] B. Shulman, A. Sharma, and D. Cosley, "Predictability of popularity: Gaps between prediction and understanding," *Tenth International AAAI Conference on Web and Social Media*, 2016.

[12] T. K. Zekarias, S. Nasrullah, B. Leila, S. Amira, M. Alberto, and G. Sarunas, "Cas2vec: Network-agnostic cascade prediction in online social networks," in *The Fifth International Conference on Social Networks Analysis, Management and Security(SNAMS-2018)*, ser. SNAMS, 2018.

[13] W. Hu, K. K. Singh, F. Xiao, J. Han, C.-N. Chuah, and Y. J. Lee, "Who will share my image?: Predicting the content diffusion path in online social networks," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 252–260.

[14] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei, "Sequential prediction of social media popularity with deep temporal context networks," *arXiv preprint arXiv:1712.04443*, 2017.

[15] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[16] L. De Vries, S. Gensler, and P. S. Leeflang, "Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing," *Journal of Interactive Marketing*, vol. 26, no. 2, 2012.

[17] I. P. Cvijikj and F. Michahelles, "Online engagement factors on facebook brand pages," *Social Network Analysis and Mining*, vol. 3, 2013.

[18] S. Mohammadi, R. Farahbakhsh, and N. Crespi, "Popularity evolution of professional users on facebook," *ACM ICC*, 2017.

[19] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," *Journal of Internet Services and Applications*, vol. 5, 2014.

[20] F. P. Barclay, C. Pichandy, A. Venkat, and S. Sudhakaran, "India 2014: Facebook 'like'as a predictor of election outcomes," *Asian Journal of Political Science*, vol. 23, no. ahead-of-print, pp. 1–27, 2015.

[21] R. Leung, M. Schuckert, and E. Yeung, *Attracting user social media engagement: A study of three budget airlines Facebook pages*. Springer, 2013.

[22] S. Jayasingh and R. Venkatesh, "Customer engagement factors in facebook brand pages," *Asian Social Science*, vol. 11, no. 26, 2015.

[23] A. Lombardi, "Social media consumer engagement: A study on the most popular fashion brands' fan pages," *Department of Business and Management, LUISS Guido Carli*, 2012.

[24] M. Pronschinske, M. D. Groza, and M. Walker, "Attracting facebook'fans': The importance of authenticity and engagement as a social networking strategy for professional sport teams," *Sport marketing quarterly*, vol. 21, no. 4, p. 221, 2012.

[25] S. Bhattacharyya, S. Banerjee, and I. Bose, "Predicting online reviewer popularity: A comparative analysis of machine learning techniques," in *Workshop on E-Business*. Springer, 2016, pp. 22–28.

[26] E. Ferrara, R. Interdonato, and A. Tagarelli, "Online popularity and topical interests through the lens of instagram," in *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 2014, pp. 24–34.

[27] Y. A. Díaz-Beristain, N. Cruz-Ramírez *et al.*, "Retweet influence on user popularity over time: An empirical study," in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2016, pp. 38–48.

[28] I. Arapakis, M. Lalmas, B. B. Cambazoglu, M.-C. Marcos, and J. M. Jose, "User engagement in online news: Under the scope of sentiment, interest, affect, and gaze," *Journal of the Association for Information Science and Technology*, vol. 65, no. 10, pp. 1988–2005, 2014.

[29] A. Susarla, J.-H. Oh, and Y. Tan, "Social networks and the diffusion of user-generated content: Evidence from youtube," *Information Systems Research*, vol. 23, no. 1, pp. 23–41, 2012.

[30] S. Bakhshi, D. A. Shamma, and E. Gilbert, "Faces engage us: Photos with faces attract more likes and comments on instagram," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 965–974.

[31] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 621–630.

[32] F. Figueiredo, M. A. Gonçalves, and J. M. Almeida, "Improving the effectiveness of content popularity prediction methods using time series trends," *arXiv preprint arXiv:1408.7094*, 2014.

[33] C. Hu, Y. Hu, W. Xu, P. Shi, and S. Fu, *Understanding Popularity Evolution Patterns of Hot Topics Based on Time Series Features*. Cham: Springer International Publishing, 2014, pp. 58–68. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11119-3_6

[34] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang, "Image popularity prediction in social media using sentiment and context features," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 907–910.

[35] K. Yamaguchi, T. L. Berg, and L. E. Ortiz, "Chic or social: Visual popularity analysis in online fashion networks," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 773–776.

[36] K. Swani, G. R. Milne, B. P. Brown, A. G. Assaf, and N. Donthu, "What messages to post? evaluating the popularity of social media communications in business versus consumer markets," *Industrial Marketing Management*, vol. 62, pp. 77–87, 2017.

[37] P. Gong and H. Wu, "A cache partition policy of ccn based on content popularity," *International Journal of Advanced Science and Technology*, vol. 92, pp. 9–16, 2016.

[38] Z. Wang, W. Zhu, M. Chen, L. Sun, and S. Yang, "Cpcdn: Content delivery powered by context and user intelligence," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 92–103, 2015.

[39] P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng, "Popularity prediction in microblogging network: a case study on sina weibo," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 177–178.

[40] S. Elsharkawy, G. Hassan, T. Nabhan, and M. Roushdy, "Towards feature selection for cascade growth prediction on twitter," in *Proceedings of the 10th International Conference on Informatics and Systems*. ACM, 2016, pp. 166–172.

[41] Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang, "Retweet prediction with attention-based deep neural network," in *CIKM*, 2016.

[42] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in twitter." in *ICWSM*, 2011.

[43] J. Hessel, L. Lee, and D. Mimno, "Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 927–936.

[44] Y. Duan, X. Wang, Y. Yang, Z. Huang, N. Xie, and H. T. Shen, "Poi popularity prediction via hierarchical fusion of multiple social clues," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 1001–1004.

[45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013.

[46] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1235–1244.

[47] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

[48] P. Turney, "Mining the web for synonyms: Pmi-ir versus lsa on toefl," *Machine Learning: ECML 2001*, pp. 491–502, 2001.

[49] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.

[50] A. Islam and D. Inkpen, "Second order co-occurrence pmi for determining the semantic similarity of words," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2006, pp. 1033–1038.

[51] G. Recchia and M. N. Jones, "More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis," *Behavior research methods*, vol. 41, no. 3, pp. 647–656, 2009.

[52] T. Martin, "community2vec: Vector representations of online communities encode semantic relationships," in *Proceedings of the Second Workshop on NLP and Computational Social Science*, 2017, pp. 27–31.

[53] D. Liang, J. Altosaar, L. Charlin, and D. M. Blei, "Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence," in *Proceedings of the 10th ACM conference on recommender systems*. ACM, 2016, pp. 59–66.

[54] M. Kaminskas and D. Bridge, "Measuring surprise in recommender systems," in *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design (Workshop Programme of the 8th ACM Conference on Recommender Systems)*, 2014.

[55] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 2012, pp. 298–307.

[56] E. Spertus, M. Sahami, and O. Buyukkokten, "Evaluating similarity measures: a large-scale study in the orkut social network," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 678–684.

[57] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 151–160.

[58] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Facta: a text search engine for finding associated biomedical concepts," *Bioinformatics*, vol. 24, no. 21, pp. 2559–2560, 2008.

[59] C. G. Akcora, B. Carminati, and E. Ferrari, "Network and profile based measures for user similarities on social networks," in *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*. IEEE, 2011, pp. 292–298.

[60] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[61] O. Barkan and N. Koenigstein, "Item2vec: Neural item embedding for collaborative filtering," *arXiv preprint arXiv:1603.04259*, 2016.

[62] N. Hollis, "The value of a social media fan," *Millward Brown*, 2011.

[63] D. L. Hoffman and M. Fodor, "Can you measure the roi of your social media marketing?" *MIT Sloan Management Review*, vol. 52, no. 1, 2010.

[64] B. Yu, M. Chen, and L. Kwok, "Toward predicting popularity of social marketing messages," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2011, pp. 317–324.

[65] Facebook, "What are global pages?" 2014. [Online]. Available: https://www.facebook.com/business/help/331800410323820

[66] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, ser. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[67] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 177–186.

[68] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 1855–1870.

[69] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, 2013.

[70] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[71] S. Hong and D. Nadler, "Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience," *Government Information Quarterly*, vol. 29, no. 4, pp. 455–461, 2012.

[72] Y. Yu and X. Wang, "World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets," *Computers in Human Behavior*, vol. 48, pp. 392–400, 2015.

[73] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Record*, vol. 42, no. 2, pp. 17–28, 2013.

[74] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing social networks*. SAGE Publications Limited, 2013.

[75] M. Cha, A. Mislove, and K. P. Gummadi, "A Measurement-driven Analysis of Information Propagation in the Flickr Social Network," in *In Proceedings of the 18th International World Wide Web Conference (WWW'09)*, Madrid, Spain, April 2009.

[76] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[77] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[78] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

Thesis Publications

- S. Mohammadi, R. Farahbakhsh, and N. Crespi. *Popularity evolution of professional users on facebook.*, IEEE International Conference on Communications (ICC). 2017, IEEE.

- S. Mohammadi, R. Farahbakhsh, and N. Crespi. *User Reactions Prediction Using Embedding Features*, (Globecom), IEEE Global Communications Conference. 2018.

- S. Mohammadi, R. Farahbakhsh, and N. Crespi. *Who Will Like the Post? A Case Study of Predicting Likers on Flickr*, Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2018, IEEE.

- S. Mohammadi, L. Jie-Yu, R. Farahbakhsh, and N. Crespi. *A Data-driven Study on Long-Term Evolution of Professional Users? Popularity on Facebook.*, Submitted - IEEE Access

- R. Farahbakhsh, S. Mohammadi, X. Han, A. Cuevas, and N. Crespi. *Evolution of publicly disclosed information in Facebook profiles*, In Trustcom/BigDataSE/ICESS, 2017 IEEE (pp. 9-16).

- P. Rajapaksha, R. Farahbakhsh, S. Mohammadi, M. N. Dailey, and N. Crespi. *Video Content Delivery Enhancement in CDNs Based on Users' Social Information.*, Globecom Workshops (GC Wkshps), 2016 IEEE.