User Identity Linkage across Online Social Networks

Comprehensive Report Presentation

PhD Advisor:
Prof. Ponnurangam Kumaraguru

Rishabh Kaushal
PhD15008

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY
DELHI

Evaluation Committee

Thanks to the Committee Members

External Reviewer:

Dr. Alpana Dubey

Internal Reviewers:

Dr. Rajiv Ratn Shah

Dr. V. Raghava Mutharaju

Faculty Advisor:

Dr. Ponnurangam Kumaraguru (PK)

Who Am I ?

Sponsored PhD Candidate at Precog Research Group at IIITD

Completed 4 yrs in PhD

- Working as faculty at IGDTUW
- Worked in Software Industry, 3 yrs

MS by Research (CSE), IIIT, Hyderabad

BTech (CSE), GGSIPU, Delhi

User Identity Linkage across OSNs

Part 1: Introduction

Part 2: Problem formulations

Part 3: Data Collection Approaches

Part 4: Machine Learning Approaches

Part 5: Representation Learning Approaches

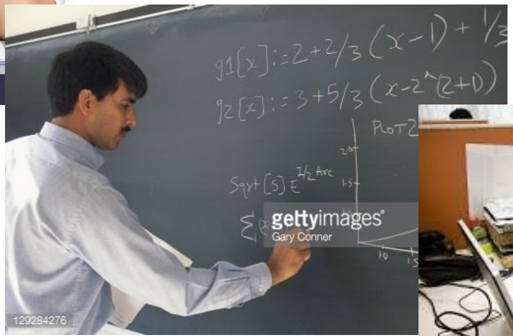
Part 6: Applications and Future Directions

Identity in Real World

Identity

Real World

Student



Teacher

Software Engineer

Father

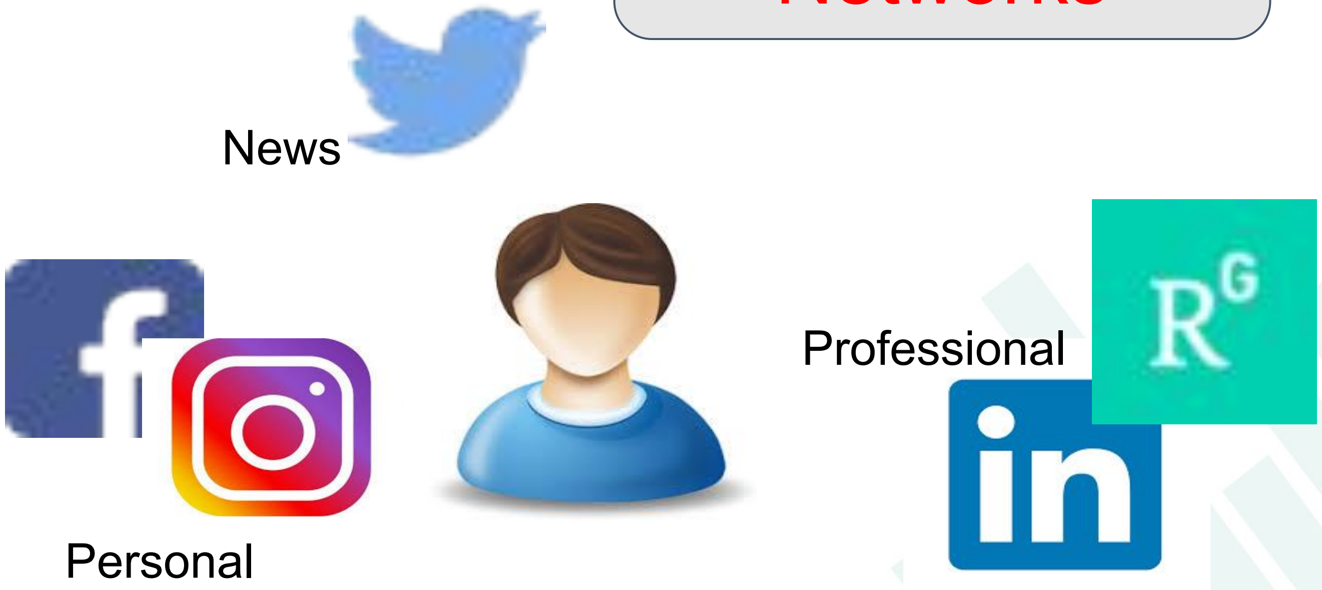
Identity in Social World

Identity has three dimensions - profile, content, and network

User joins multiple social networks

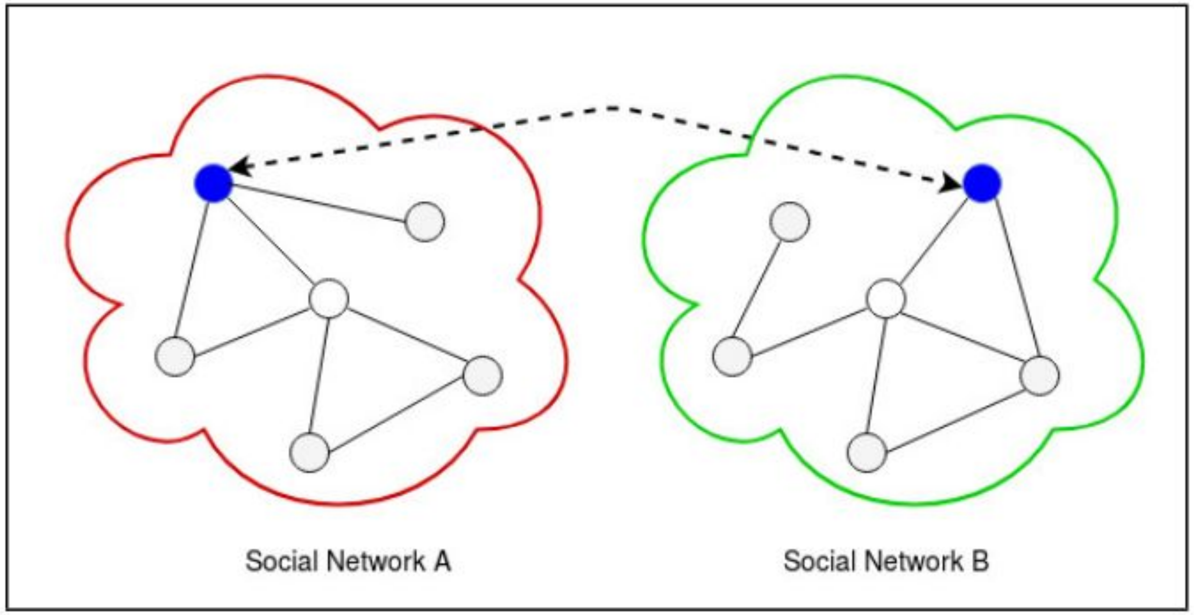
Identity

World of Social Networks



Problem: User Identity Linkage (UIL)

The goal is to find user identity on target OSN (social network B) when that user's identity is known on source OSN (social network A)



Motivation to solve UIL

User Behavior Studies

Recommendation

Influence Maximization

Audience Segregation

Privacy Disclosure

Most of the problems in social networks have traditionally been solved using user's information (behavior) in single network

With UIL, a comprehensive user information can be used to solve conventional problems with more user data

Problem Nomenclature & Formulations

In the literature UIL problem is referred by multiple names

Social Identity Linkage,

User Identity Resolution,

Social Network Reconciliation,

User Account Linkage Inference,

Profile Linkage, and

Anchor Link prediction

Two main problem formulations:-

(1) Identity Linkage

(2) Linked Identity Extraction

F1: Identity Linkage

Looking it as a conventional *classification problem* in machine learning settings

Definition 2.1 *Given two user identities I_a and I_b on OSNs a and b, respectively, the goal is to learn a function F, which predicts whether I_a and I_b belong to the same individual or not.*

$$F(I_a, I_b) = \begin{cases} 1, & \text{if } I_a \textit{ and } I_b \textit{ belong to the same user} \\ 0, & \text{otherwise} \end{cases}$$

Taken from comprehensive report

F1: Identity Linkage (methodology)

1. We consider all possible identity pairs $< I_a, I_b >$ comprising of identities belonging to two social networks a and b as part of the input dataset D.
2. Each identity pair $< I_a, I_b >$ has a *label* associated with it, whose value is binary, either 1 or 0, indicating whether two identities I_a and I_b on OSNs a and b, belong to the same or different individuals, respectively.
3. We split the dataset D into training and test datasets. We use the label as supervisory information for learning of the function F. Evaluation is done based on standard metrics, as discussed in Table 1.

F1: Identity Linkage (evaluation)

Evaluation Metric	Interpretation in context of UIL problem
True Positive (TP)	User identities I_a and I_b belong to the same person and the learned function F also predicts the same person.
True Negative (TN)	User identities I_a and I_b do not belong to the same person and the learned function F also says they do not belong to the same person.
False positive (FP)	User identities I_a and I_b do not belong to the same person but the learned function F says they belong to the same person.
False negative (FN)	User identities I_a and I_b belong to the same person but the learned function F says they do not belong to the same person.

F2: Linked Identity Extractor

The goal is to learn a *ranking function* which given a single user identity on one social network (source), orders the identities on another social network (target) such that correct linked identity appears among the top-k identities extracted from the target network.

Definition 2.2 *Given a user identity I_a on source OSN_a , the goal is to learn a function F_{rank} that finds top-k user identities $< I_b^1, I_b^2, \ldots, I_b^k >$, one out of which is likely to belong to the same individual whose identity I_a on OSN_a is already known.*

F2: Linked Identity Extractor (methodology)

1. We consider all possible identity pairs $< I_a, I_b >$ comprising of identities belonging to the two social networks a and b to be part of input dataset D.

2. For each user identity I_a in linked identity pair $< I_a, I_b >$, using different ranking functions, we find an ordered list of identities $< I_b^1, I_b^2, ..., I_b^k >$.

3. Subsequently, we perform evaluation on the basis of metrics discussed in Table 2.

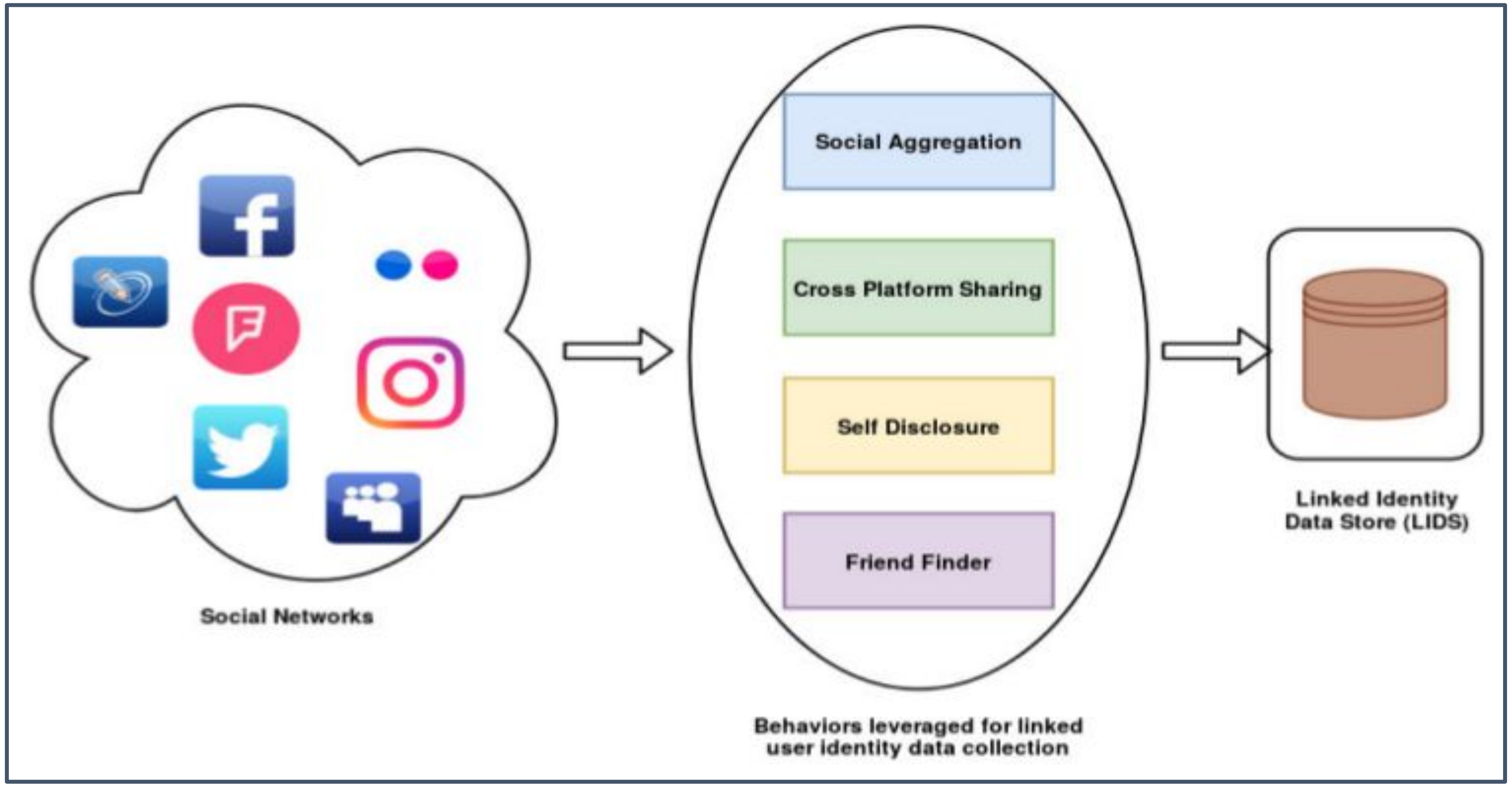
F2: Linked Identity Extractor (evaluation)

Evaluation Metric	Interpretation in the context of UIL problem
Success/Hit at Rank k (S@k)	The proportion of times that correct linked identity I_b is present among the top-k identities that we retrieve.
Mean Reciprocal Rank (MRR)	The average rank at which the linked identity I_b occurs in the top-k identities that we retrieve.

Table 2: Explanation of evaluation metric in the context of user identity linkage

Data Collection

First step is to collect ground truth, user identities on different social networks that belong to the same individual



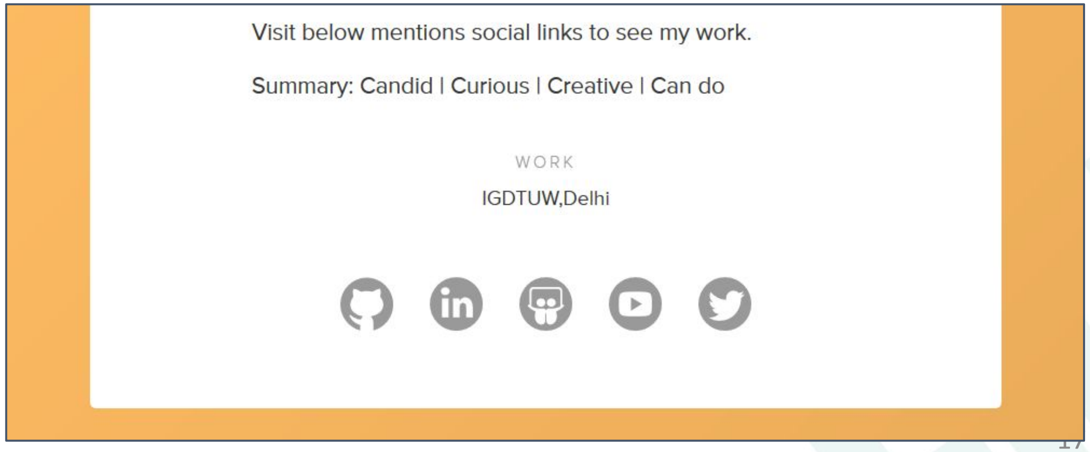
Social Aggregation

There are several social aggregating websites on which users create an account and provide details of their multiple social network accounts.

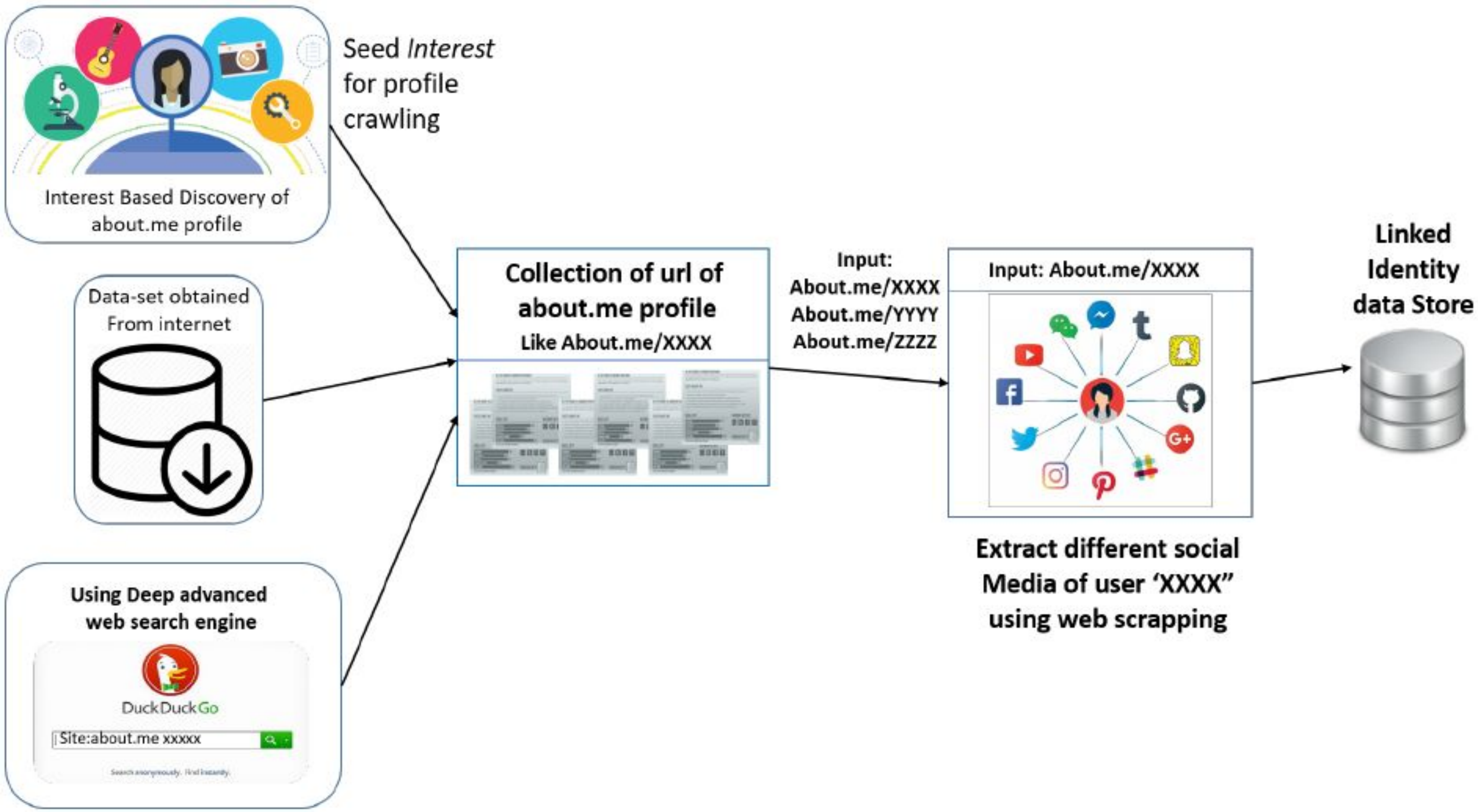
Perito et al. → Google profiles

Liu et al. → About.me profiles

....

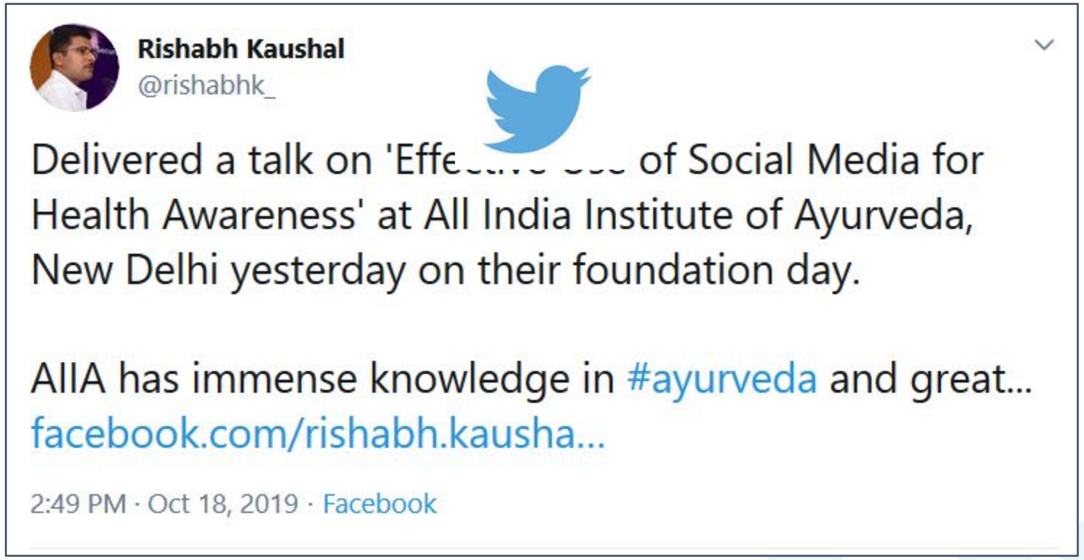


Social Aggregation

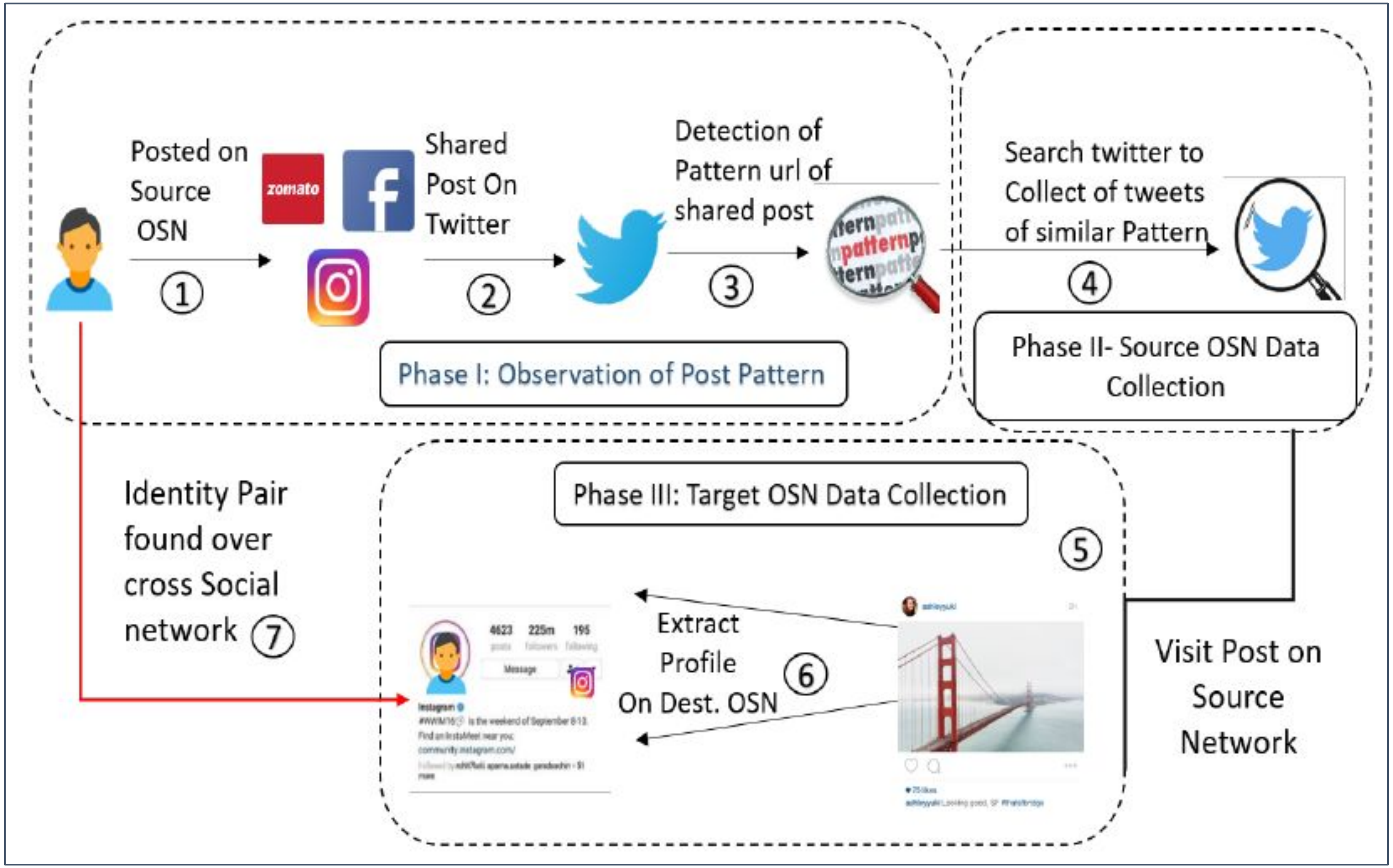


Cross Platform Sharing

- Cross platform sharing (Jain et al.) refers to user behavior in which users post the same content across multiple social network



Architecture



Self Disclosure

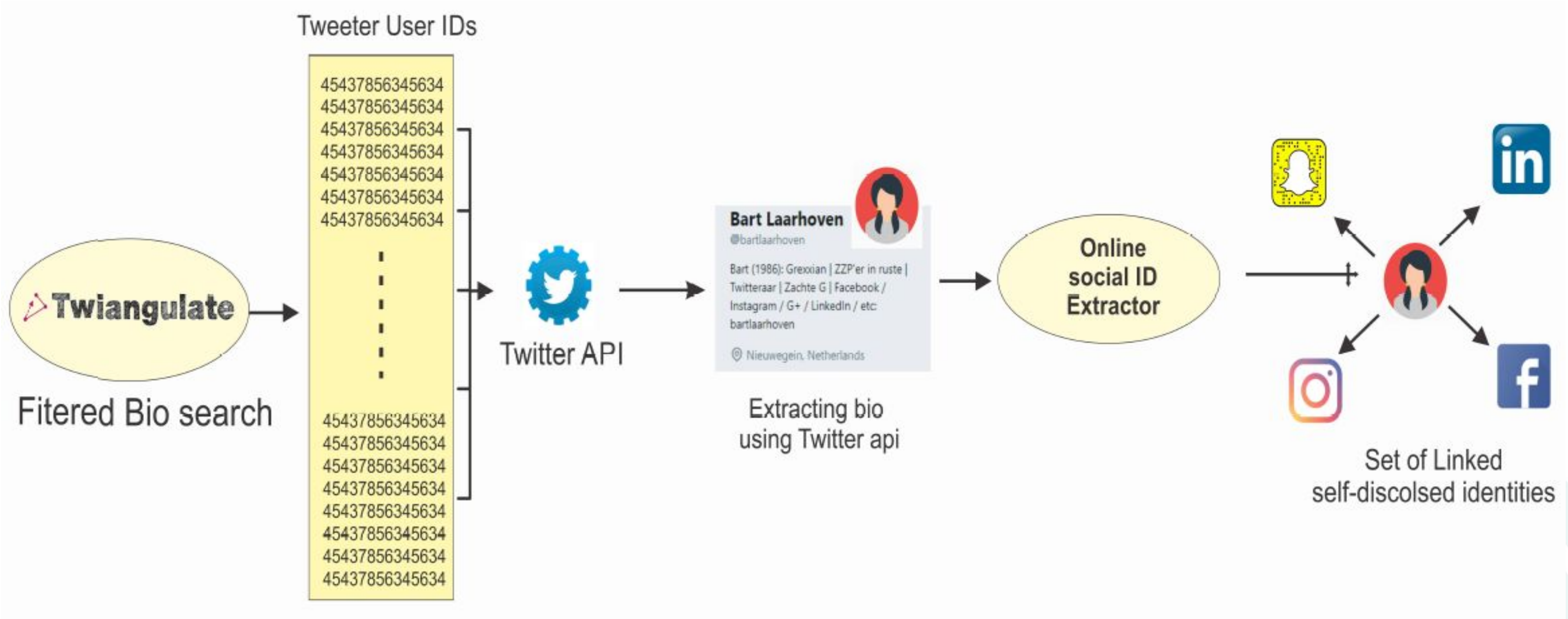
Whenever a user signs up on OSN, there is an option to provide a user description. At times, users provide details of their identities on other OSNs, which we refer to as *self disclosure*.

Example: On facebook profile page, a user can provide details of his/her Twitter handle

Kong et al. used FourSquare profile page to extract Twitter profile information.


Rishabh Kaushal
Rishabh
Home
Create
Secure
Rishabh Kaushal
Update Info 2
Activity Log 20+
Timeline
About
Friends 965
Photos
Archive
More
About
Overview
Work and Education
Places You've Lived
Assistant Professor at Indira Gandhi Delhi Technical University for Women - IGDTU and Assistant Professor at Indira Gandhi Delhi Technical University
Past: GGSIPU and Thapar University
098107 74114
rk.iiit@gmail.com
rishabhk_ (Twitter)
rk.iiit (Google Talk)

Architecture

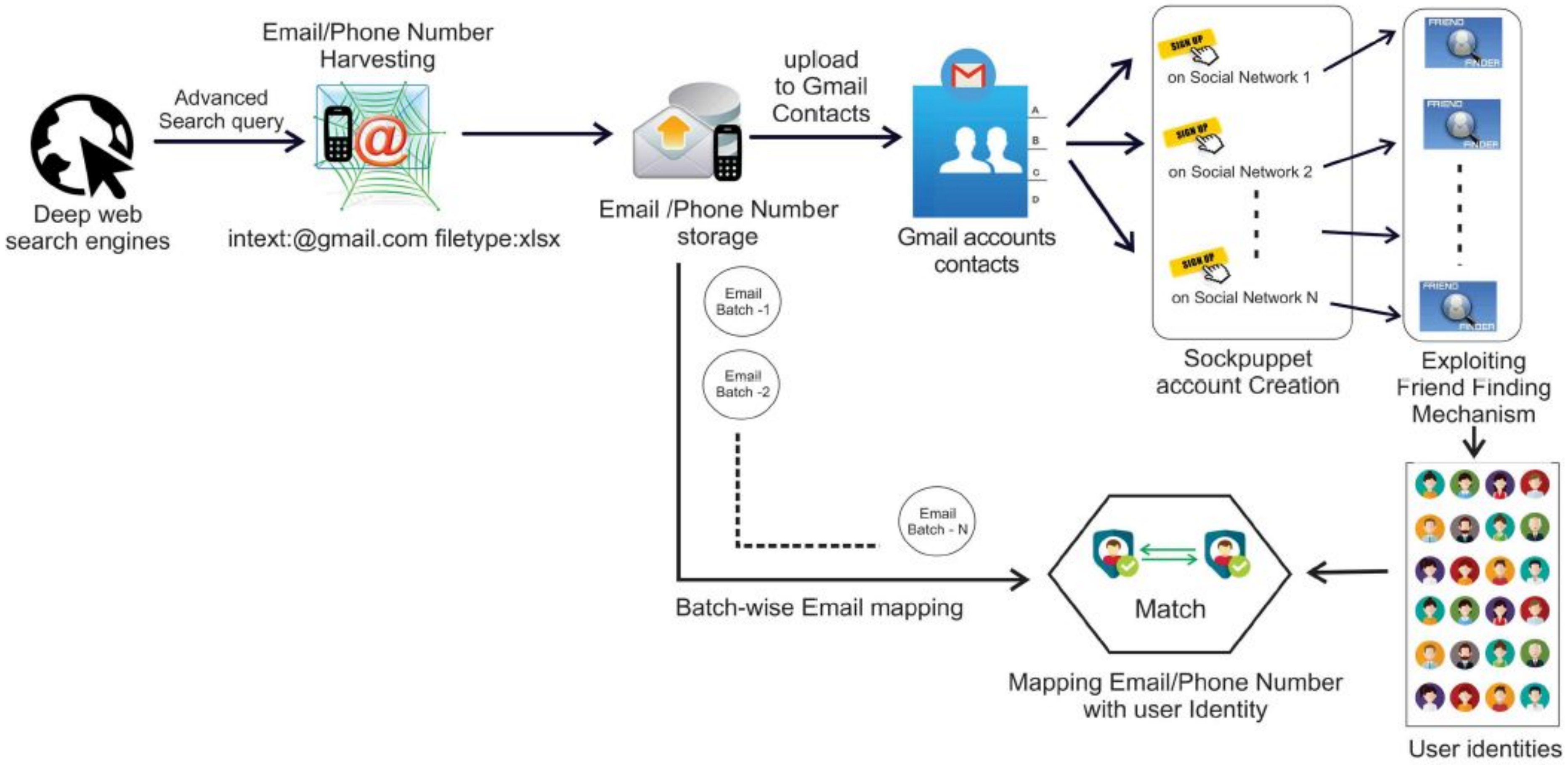


Friend Finder

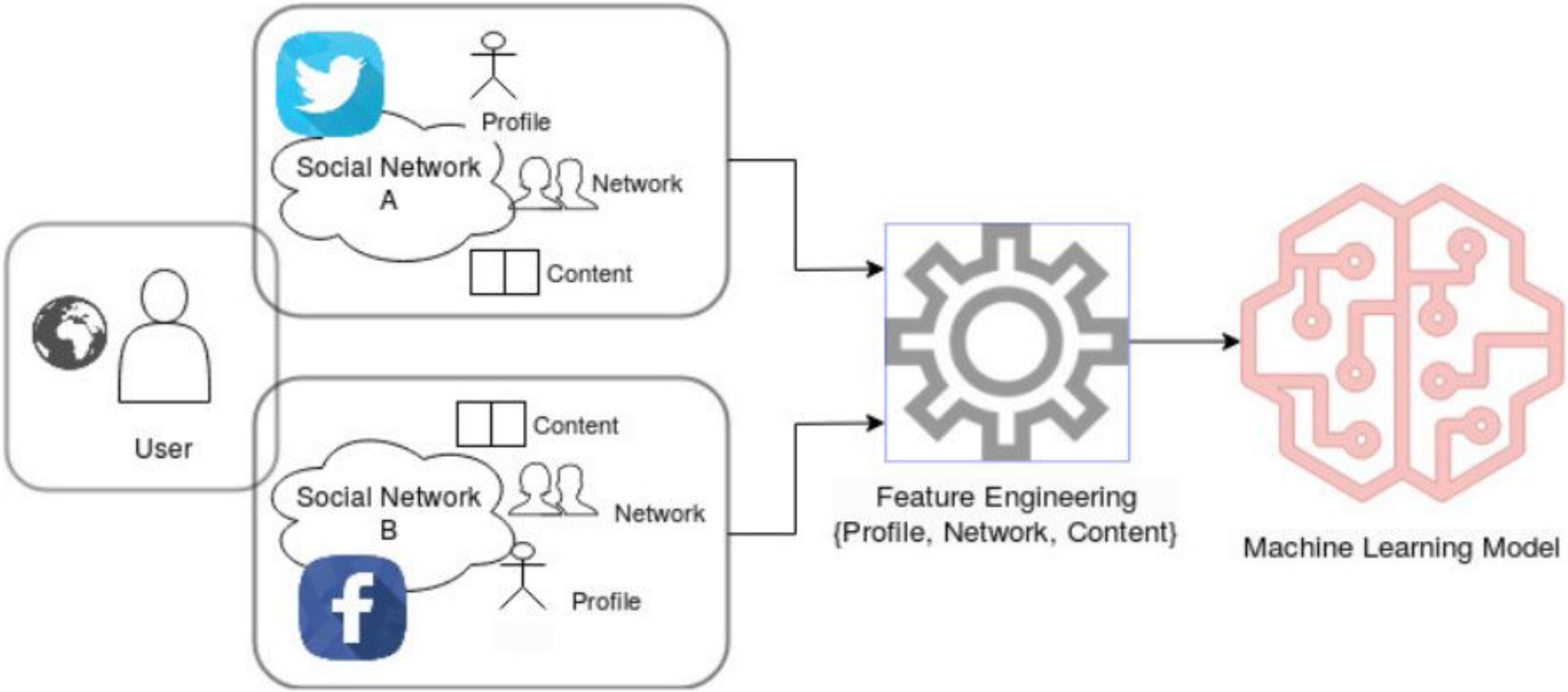
- Whenever a user joins a new OSN, we sign up using our unique identifier, say email or phone number.
- This information is used by OSN to find our friends in our email contacts or phone contacts.
- Using this information, OSN offers a friend finder option to help connect to those friends who already have an account in OSN.

  (Goga et al.)

Architecture



Machine Learning Approach



Profile Features

Profile features comprise of user's basic information like username, display name, location, and profile picture.

Perito et al. proposed to connect user identities only based on **usernames**.

Zafarani et al. proposed a framework called MOBIUS (modeling behavior for identifying users across sites) for connecting user identities across social media sites.

Content Features

In this section, we discuss prior works that derive features from the content posted by users on various OSNs.

Goga et. al. investigated three characteristic features associated with posted content, which include the **timestamp** of post, the **writing style** of the user, and the **geo-location** with the post.

Chen et al. proposed a novel STUL (spatio-temporal user linkage) model, which extracts the **spatial** and **temporal features** from the content posted by users to link user identities.

Network Features

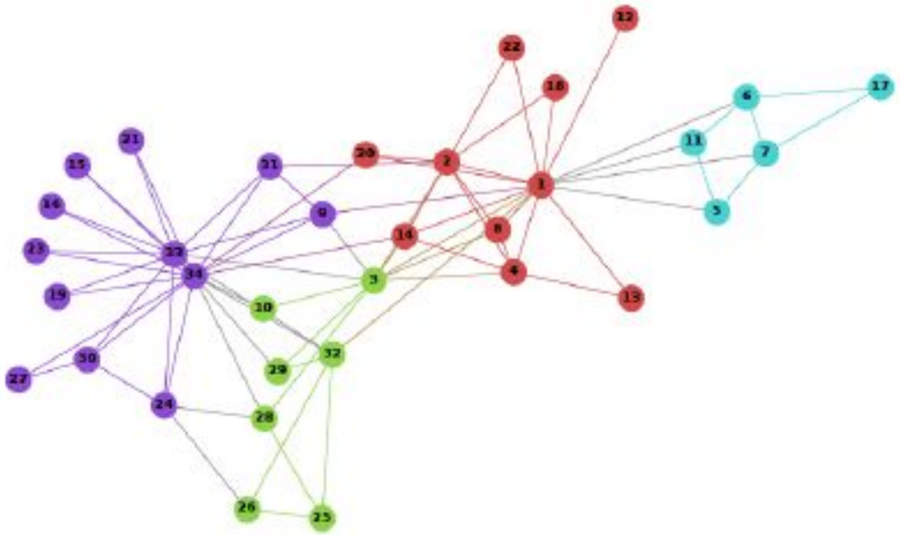
One of the fundamental principles of social networking is the concept of homophily, which implies similar users connect with each Other.

Zhou et al. proposed FRUI (Friendship Relationship Based User Identification) algorithm, which uses the fact that identical users set up **common friendship** structures in different social networks.

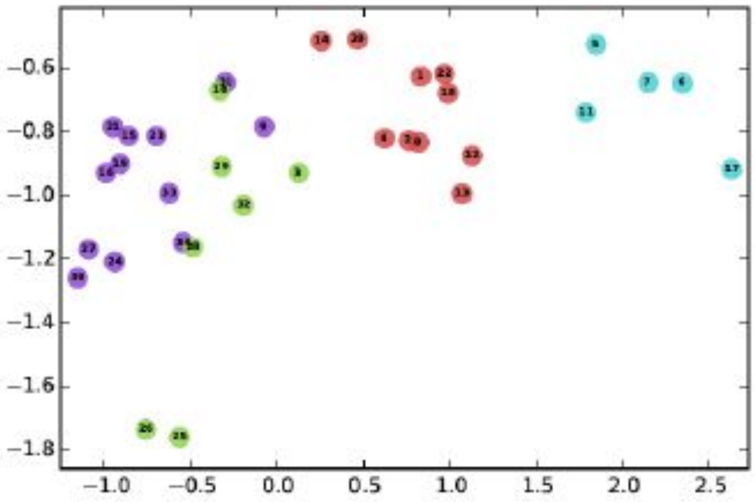
Representation Learning Approach

In the representation learning approach, features are learned implicitly rather than explicitly from profile, content, and network.

Goal: To learn *d-dimensional vectors* (embeddings) of nodes such that similar nodes in input graph have embeddings close to each other.



Input

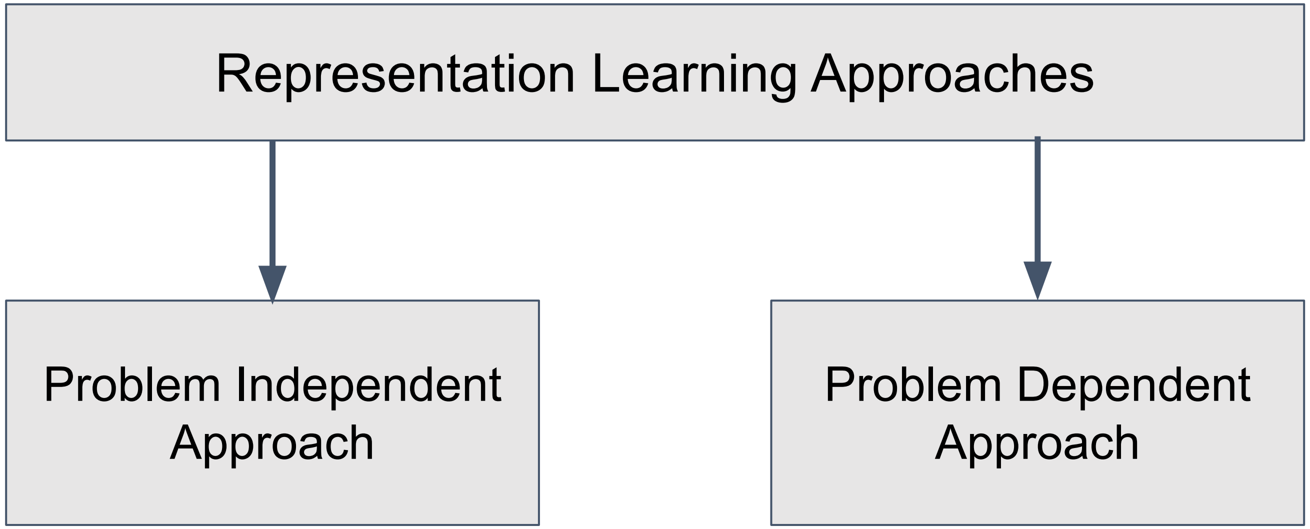


Output

Representation Learning Approach

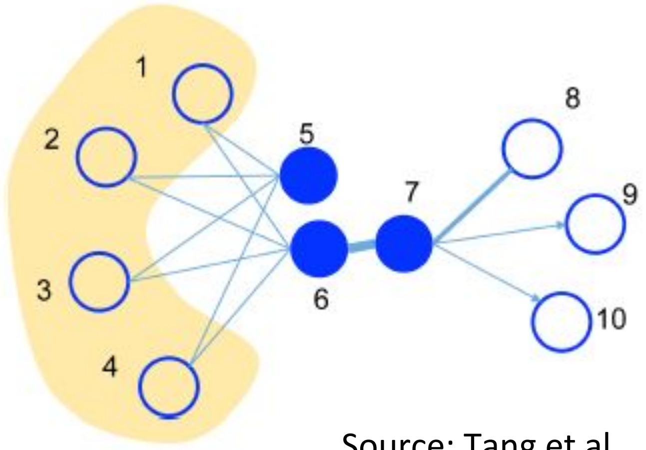
These low dimensional representations are the features learned, unlike the approach where hand-crafted features are computed explicitly.

We categorize these works into two main categories, namely, problem-independent and problem-dependent approaches.



Problem Independent Approach

Tang et al. proposed a framework, referred to as **LINE** for network embedding in large graphs.

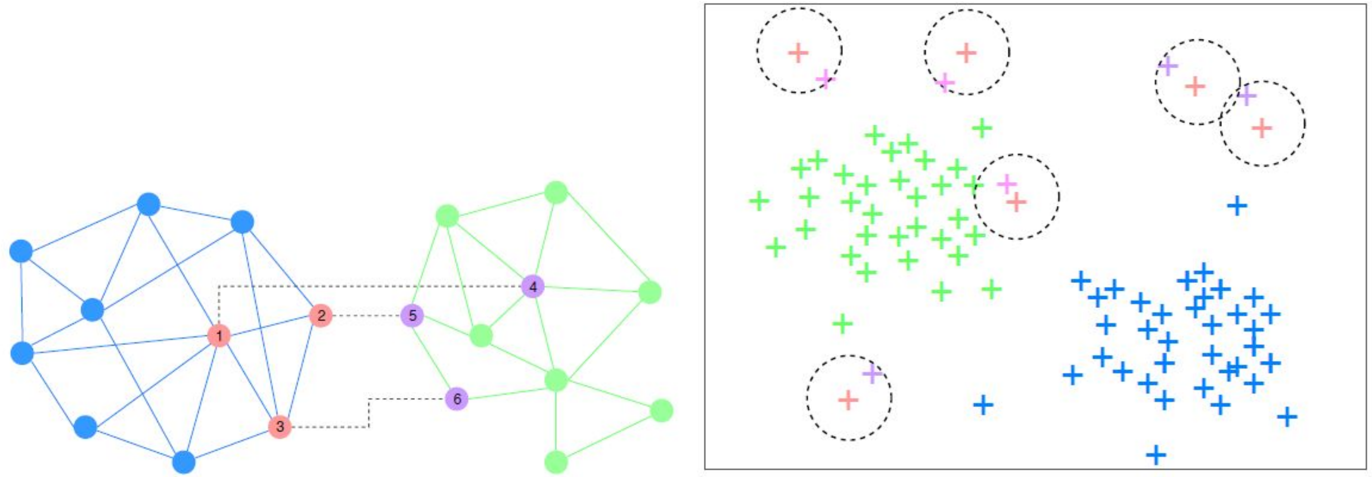


Source: Tang et al.

Perozzi et al. proposed the **DeepWalk** framework to learn node representations in a given network.

Problem Dependent Approach

In this approach, we discuss prior works that learn low-dimensional embedding focusing on the specific problem of linking user identities across social networks.



(a) Input: Two networks with cross-network linkages

(b) Output: Representation

Network based

Liu et al. proposed an Input-Output Node Embedding(IONE) framework to align user identities across social networks belonging to the same person by learning node representations that preserve **follower-followee** relationships.

Man et al. introduced a framework referred to as PALE (Predicting Anchor Links via Embedding), which predicts anchor links via embeddings. They used few known linked identities referred to as anchor links as supervisory information.

Network & Attribute based

Heimann et al. proposed the REGAL framework, which stands for representation learning-based graph alignment and is based on the cross-network **matrix factorization** method (xNetMF)

Su et al. proposed MASTER framework based on constrained dual embedding (CDE) model that simultaneously align more than two social networks and learn node embeddings at the same time.

Network & Content based

Wang et al. proposed LHNE mode referred to as linked heterogeneous network embedding model. It creates a unified framework to leverage structure and content posted by users for learning node representations.

Sajadmanestet al. proposed CRMP (Connector and Recursive Meta-Path) framework, which is a meta-path based approach. In addition to the actual friendship network, they created a content based network taking into account location, keywords, and time of the post.

Applications & Future Directions

Recommendations

Making recommendations for different aspects by using user's behavioral preferences on more than one social network is an important application.

Ozsoy et al. collected data from different online platforms, namely Twitter, BlogCatalog, Facebook, Flickr, LastFm,and YouTube to help in recommendations.

Applications & Future Directions

Link Prediction

In the context of two or more social networks, the problem of link prediction helps in finding out whether a user would join a new social network or not.

Zhang et al. proposed meta-path based approach for link prediction across multiple social networks.

Applications & Future Directions

Social Capital

Social capital of users refer to their popularity and acceptance in the social network world which prior works have measured in different ways in terms of likes, shares, engagements, and followers that users receive.

Zafarani et al. studied variations in popularity and friendship for the same users across different social networks.

Applications & Future Directions

Social Network Forensics

Malicious users perform online crimes, and while they may not leave much information on the OSN in which alleged crime was committed, but they may leave behind footprints in other OSNs.

Applications & Future Directions

User Privacy

There are privacy implications on users owing to the linkage of their identities across social networks. With online social networks, there is a collapse of user context, which has privacy implications.

Fox et al. investigated the challenges faced by professionals, particularly teachers, in managing their personal and professional identities in social media.

Applications & Future Directions

DataSet Biases

A number of data collection approaches have been used in the past to collect user identities belonging to the same user across social networks.

Each of those approaches relies on specific characteristic behaviors of users who maintain identities across multiple social networks.

Consequently, behavioral biases exhibited by users often get infested in these linked identity datasets.

Contributions & Publications

Methods for user profiling across social networks

- Comparative analysis of methods for gathering user identities belonging to the same individual across social networks

Accepted at 12th IEEE International Conference on Social Computing (SocialCom, 2019), Xiamen, China.

Contributions & Publications

Investigation of biases in identity linkage datasets

- We characterize, detect, and quantify behavioral biases in identity linkage datasets

Accepted at 35th ACM/SIGAPP Symposium on Applied Computing (SAC 2020). Brno, Czech Republic.

Contributions & Publications

NeXLink: Node embedding framework for cross-network linkages (CNLs) across social networks.

- We obtain an effective social network graph representation such that node embeddings of users belonging to CNLs are closer in embedding space than other nodes

Accepted at International School & Conference on Network Science (NetSciX, 2020), Tokyo, Japan.

International School and Conference on Network Science | Tokyo, Japan January 20-23, 2020

Contributions & Publications

Nudging Nemo: Helping Users Control Linkability Across Social Networks

- Soft interventions which alerts users whenever their behavior changes linkability of their identities across social networks

Accepted at 9th International Conference on Social Informatics (SocInfo, 2017), University of Oxford, London.

Questions ?

Thanks

Expectations

Course Work

Introduction & Literature Review

Objectives

Plan of PhD work

Publications

Future Directions

Course Work Completed

First Year

CSE 648, Privacy and Security in Online Social Media, grade B

CSE 508, Information Retrieval, grade A-

Second Year

CSE 545, Foundations of Computer Security, grade B

CSE 651, Topics in Adaptive Cyber Security, grade A-

CGPA: 8.5

Network & Content based

Xu et al. proposed two embeddings for each node that capture the structural proximity of nodes as well as the semantic similarity,

which they express in terms of common interests.

Liang et al. proposed Dynamic User and Word Embedding model (DUWE) that monitors over some time, the relationship between user and words.