Connecting Users across Social Media Sites: A Behavioral-Modeling Approach

Reza Zafarani and Huan Liu
Computer Science and Engineering, Arizona State University, Tempe, AZ
{reza@asu.edu, huan.liu@asu.edu}

Introduction and Motivation

- We can get more information about individuals by combining their partial, but complementary information on different social media sites.
- Connectivity information (who on site A is who on site B) is often not directly available for users across social media sites.
- Can we find a mapping that connects the same individual across sites?
- We believe information shared by users on social media sites provides social fingerprints of them and can be employed to identify them.

Problem Formulation

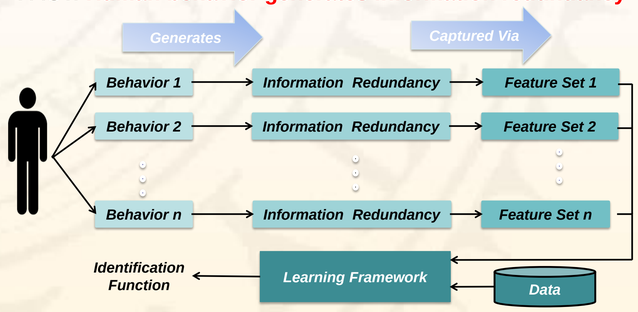
Minimum amount of Information Available on ALL Sites: Usernames

$$f(U,c) = \begin{cases} 1 & If\ c\ and\ set\ U\ belong\ to\ \mathcal{I}\ ; \\ 0 & Otherwise. \end{cases}$$

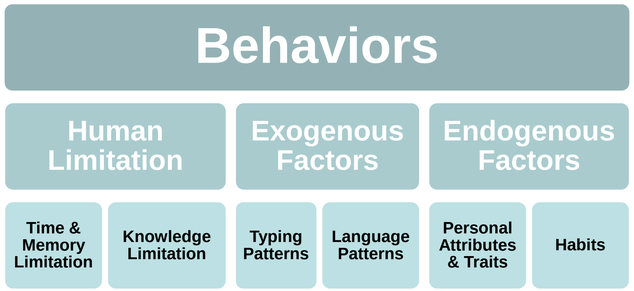
Identification Function — *Candidate Username e.g., john.doe* — *Prior Usernames e.g.,{jdoe,jdoe2013}*

MOBIUS: MOdeling Behavior for Identifying Users across Sites

FACT: Human behavior generates information redundancy



Individual Behaviors when Selecting Usernames



Human Limitations

Time and Memory Limitation

1. **Using the same usernames:** individuals use their few usernames all the time.
2. **Username Length Likelihood:** Individuals create usernames of similar length.
3. **Username Uniqueness Likelihood:** each individual has a level of uniqueness associated with her usernames.

Knowledge Limitation

1. **Limited Vocabulary:** individuals have limited vocabulary used in their usernames.
2. **Limited Alphabet Size:** individuals use a limited set of alphabet letters in their usernames and it is correlated with the language.

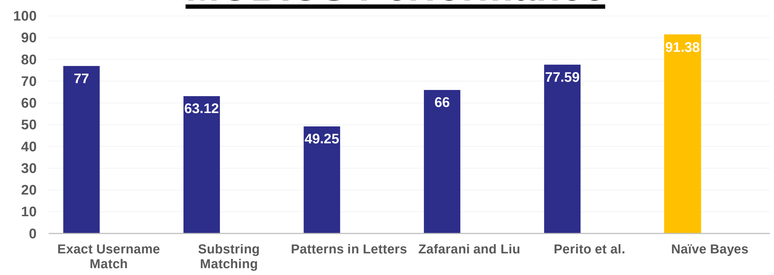
Exogenous and Endogenous Factors

- **Typing Patterns:** keyboard type impacts usernames.
- **Language Patterns:** Usernames of individuals follow a language distribution.

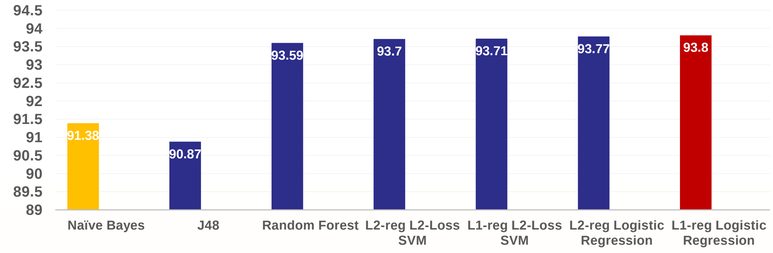
Habits:

- Modifying Previous Usernames
- Creating Similar Usernames
- Username Observation Likelihood

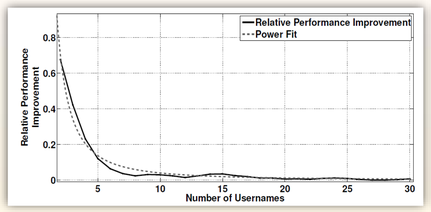
MOBIUS Performance

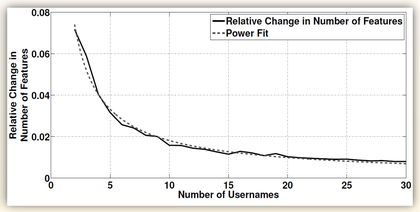


Choice of Learning Algorithm



Diminishing Returns for Adding More Usernames and More Features





This work is, in part, supported by ONR N000141110527.