UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

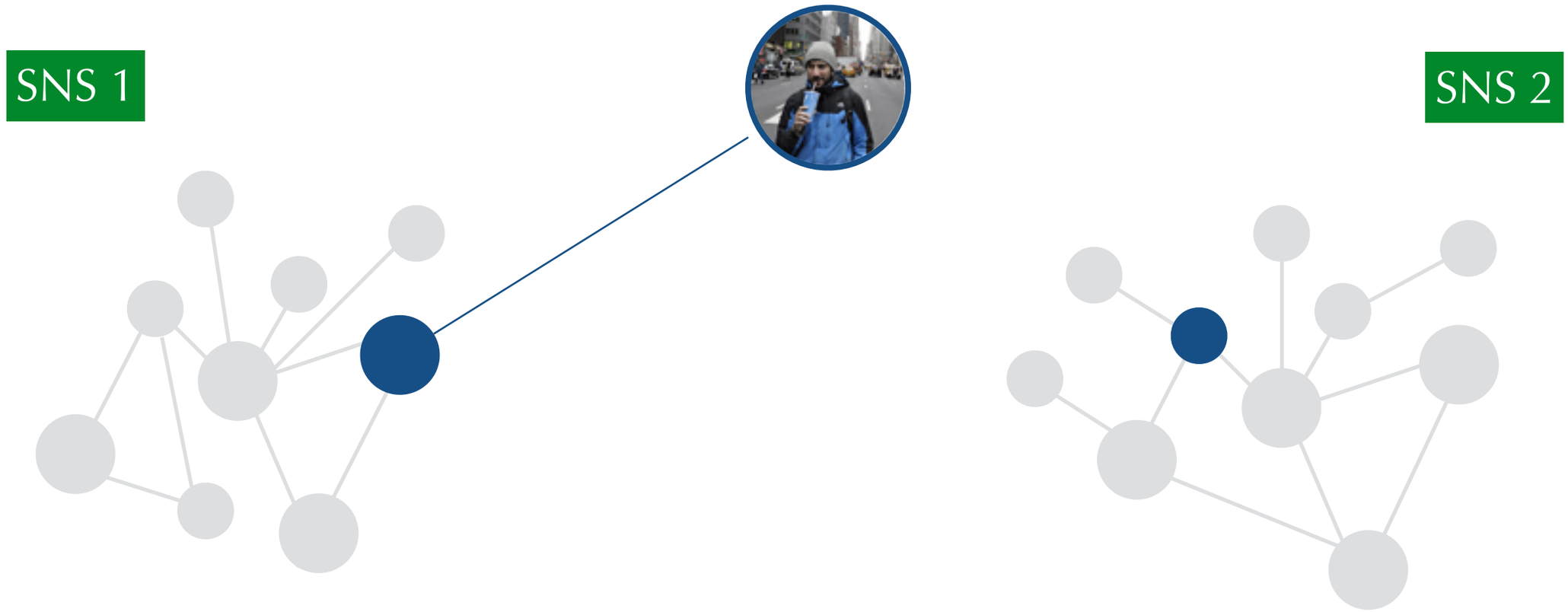
Profile matching in online social networks

Mattia Dimauro

Relatore: Sabrina Gaito
Correlatore: Matteo Zignani

Università degli Studi di Milano
Department of Computer Science
Network Protocols and Technologies Lab.

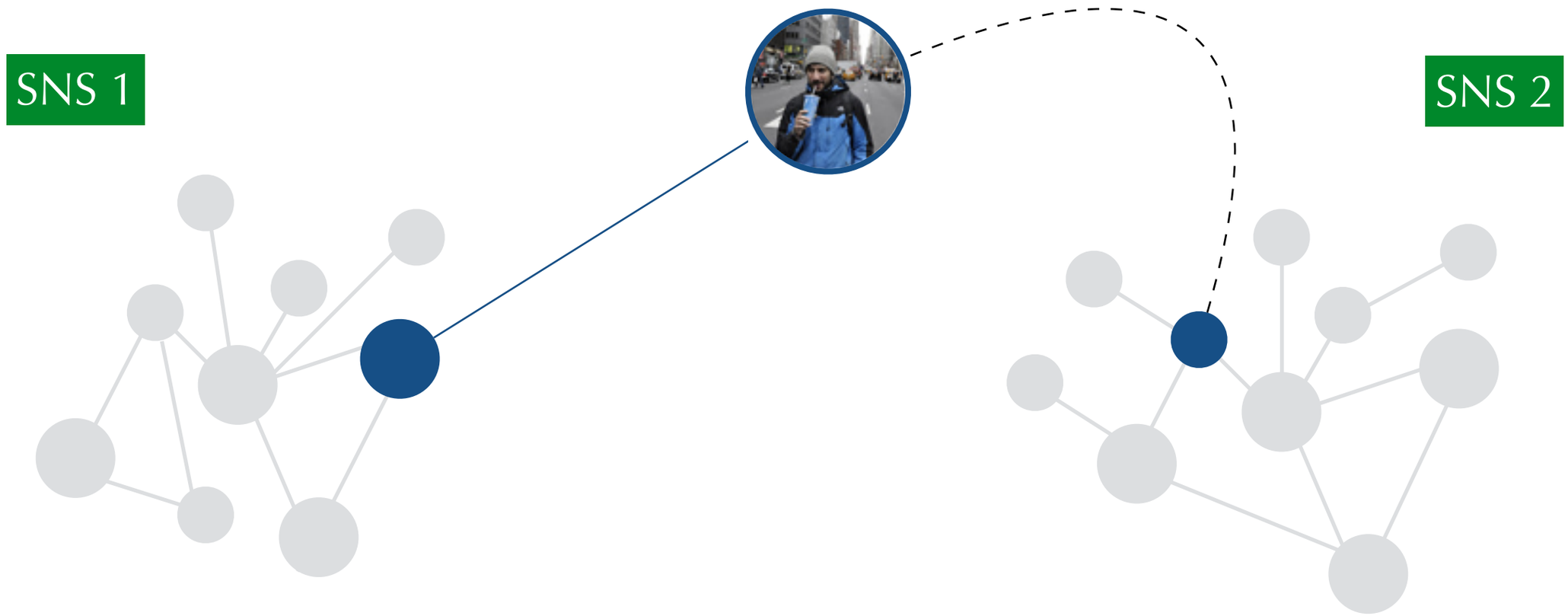
Problem | Identify an individual through different social network services



Is it possible to identify an individual through different social networks?

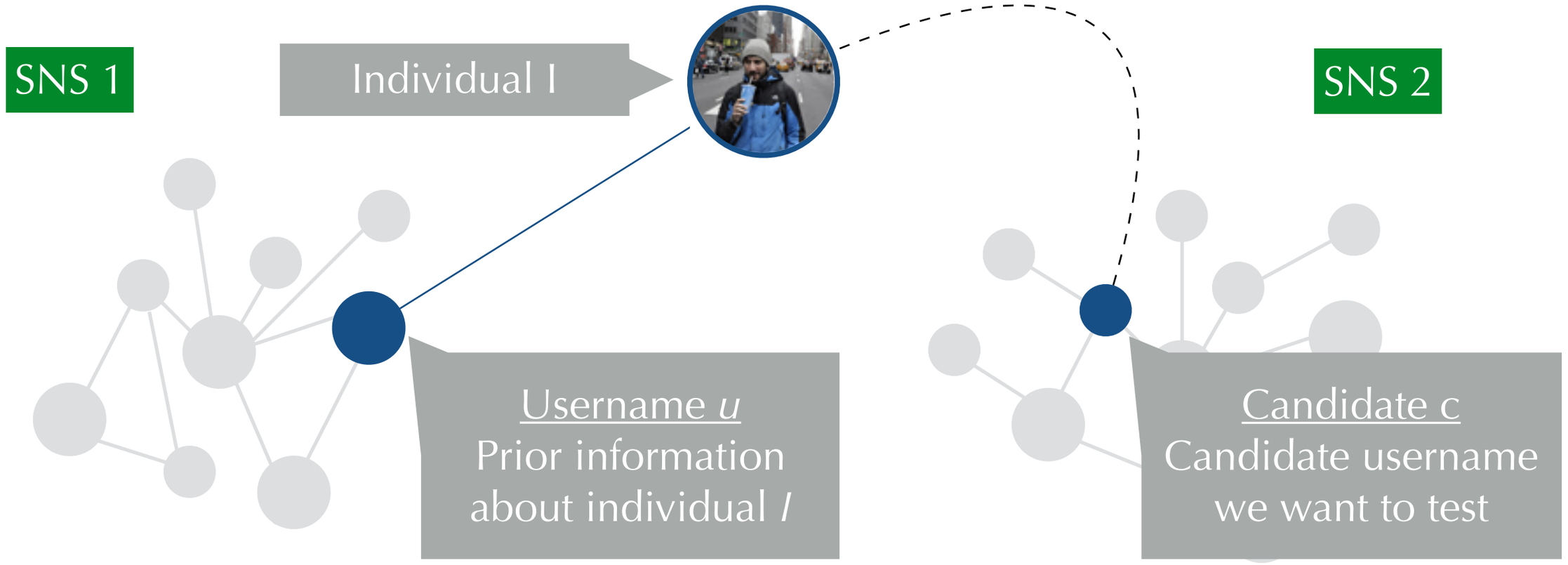
Is there a correlation between usernames on different SNS connecting real individuals?

Problem | Identify an individual through different social network services



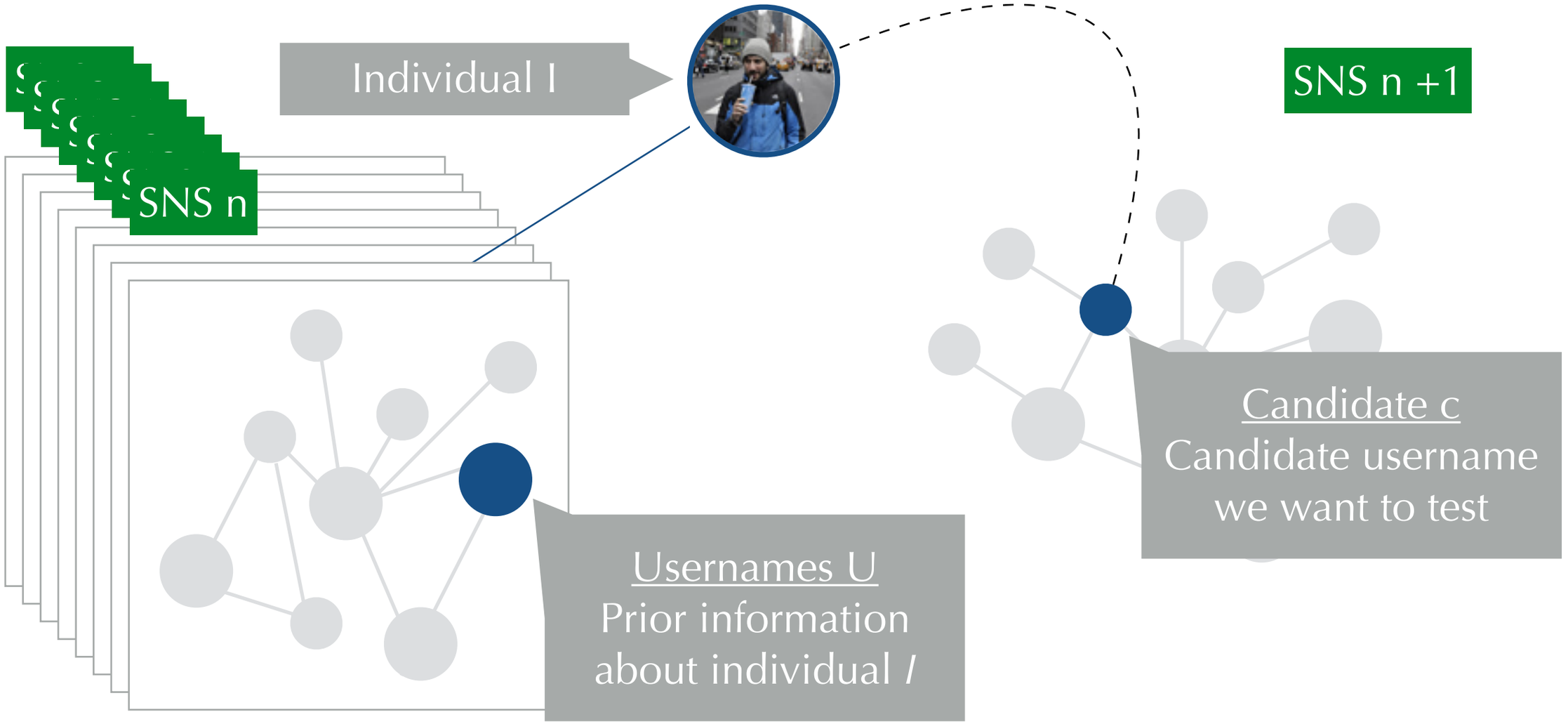
Can we find this correlation?

Problem | Identify an individual through different social network services



$$f(u,c) \begin{cases} 1 & \text{if } c \text{ and } u \text{ belongs to the same } I \\ 0 & \text{otherwise} \end{cases}$$

Problem | Identify an individual through different social network services



$$f(U,c) \begin{cases} 1 & \text{if } c \text{ and the set } U \text{ belongs to the same } I \\ 0 & \text{otherwise} \end{cases}$$

State of the art | Behavioural model

Individual often exhibit consistent **behavioural patterns** while selecting their usernames

These patterns result in **information redundancies** that help identify individuals across social media sites

Human Limitations

Limited Time and **memory**

Limited knowledge

Exogenous Factors

Typing pattern

Endogenous F.

Habits such as **abbreviation**, **prefixes**, **suffixes**

Contribution | Collecting dataset and data scraping

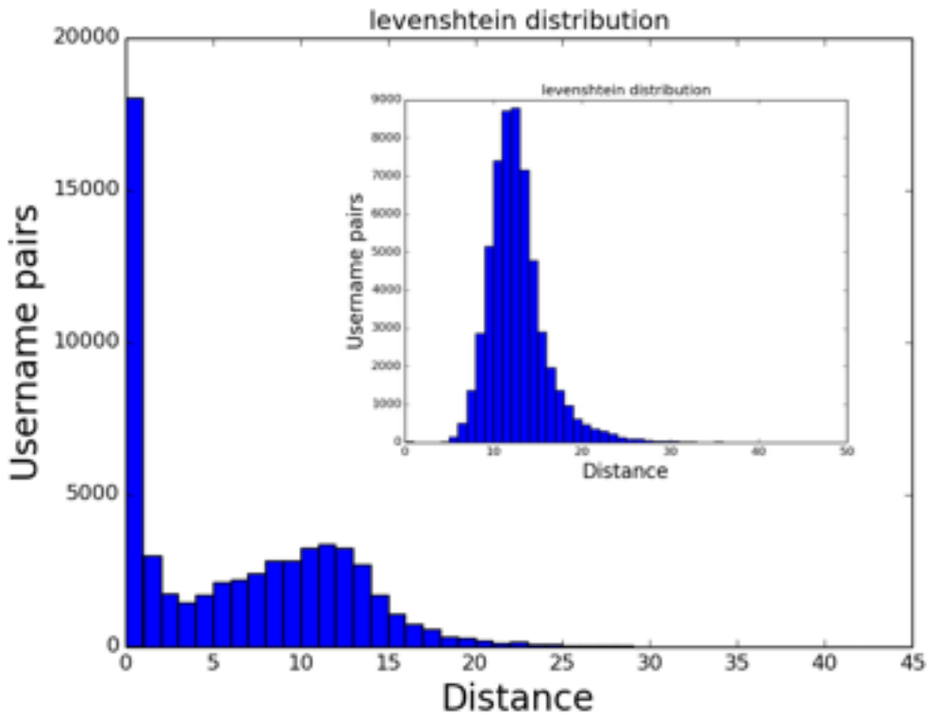
- Retrieved **15341 profiles** from *Alternion*
- Each one have ~**4,6 SNS** linked profile
- Distributed across **168 different SNS**
- **~170k usernames pairs**

Similarity distribution among usernames pair in the dataset.

Using strings similarity metrics

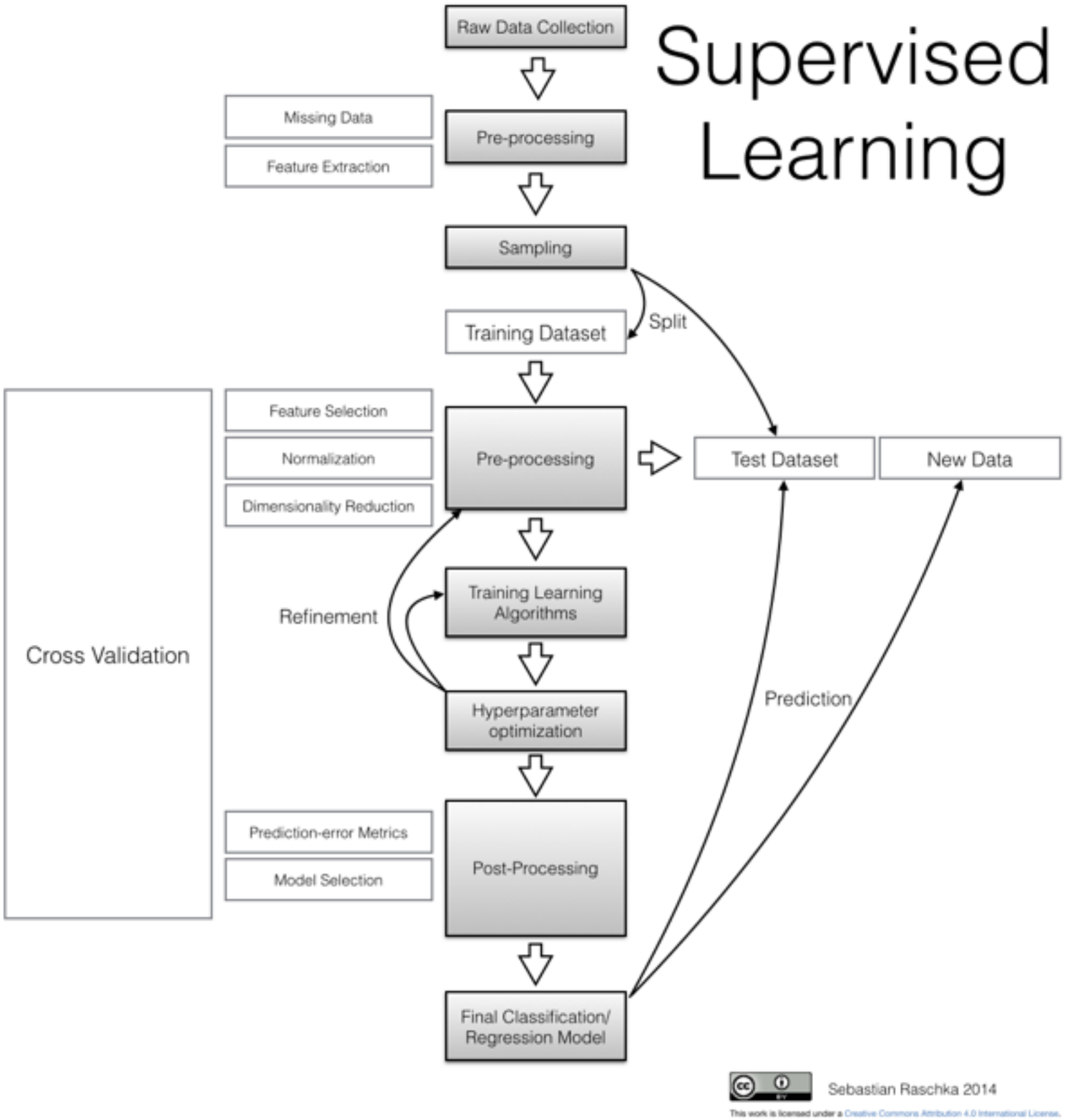
Levenshtein distance

Jaccard index

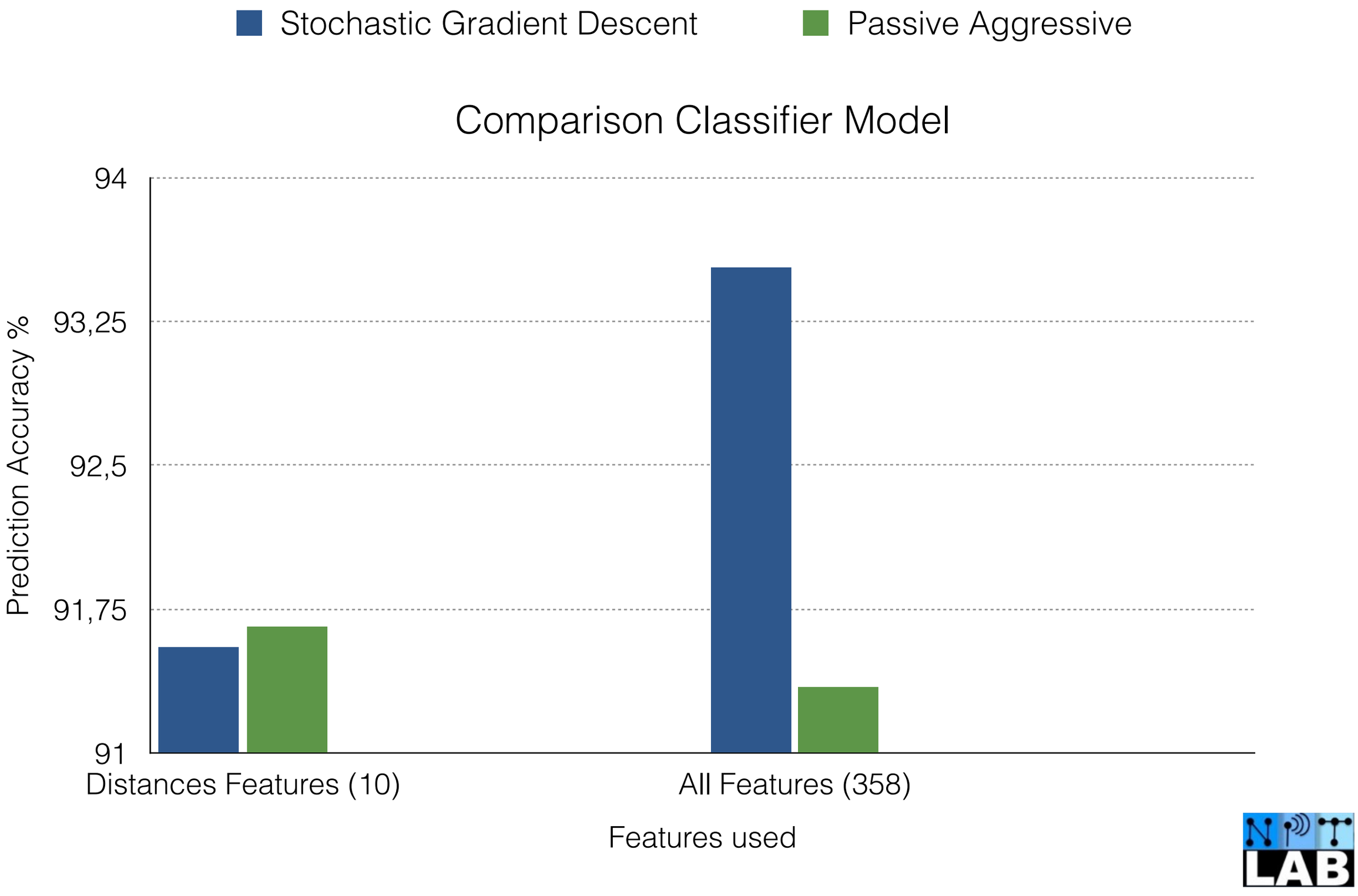


The inner plot show the distribution of random usernames pairs

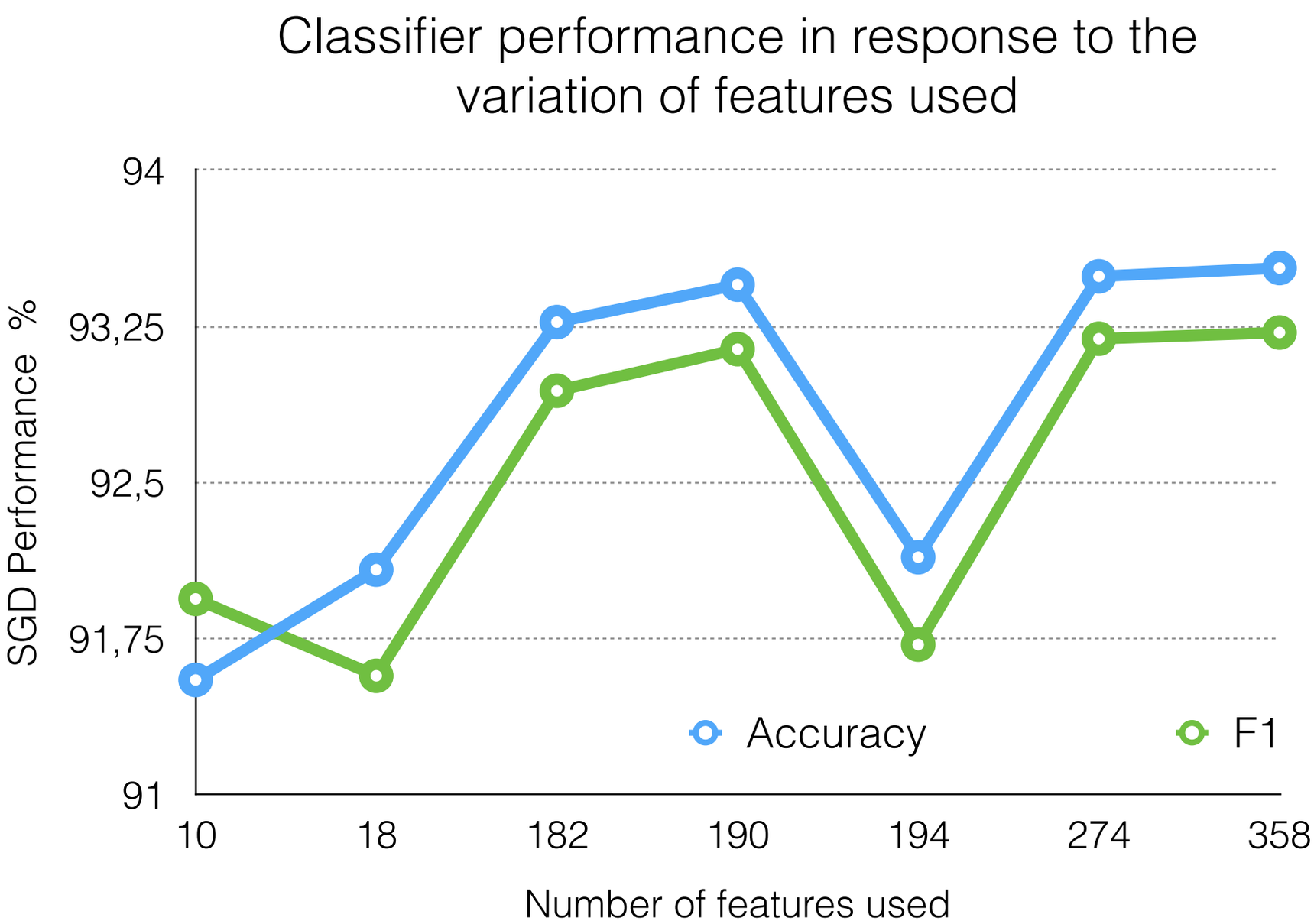
Contribution | Classification problem



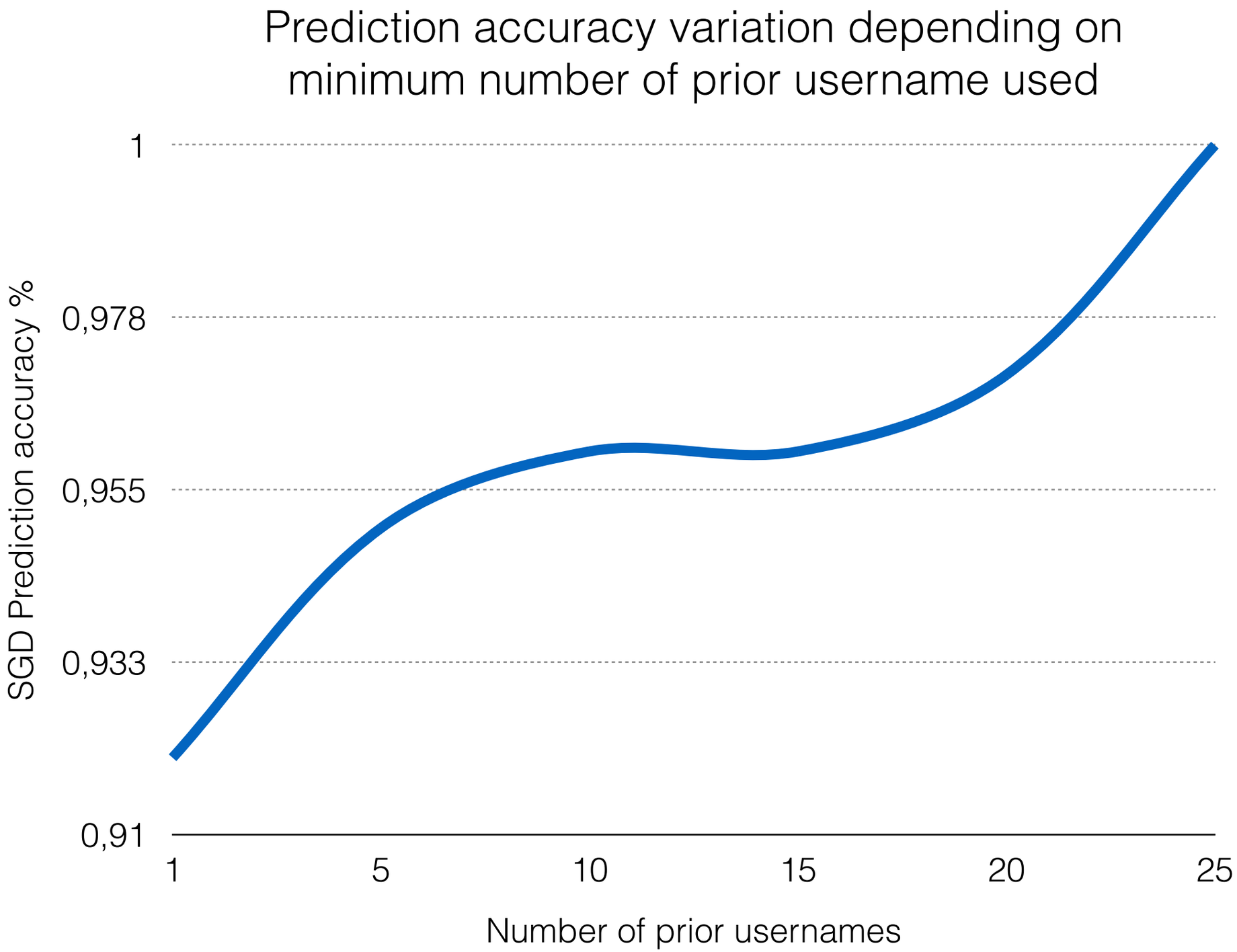
Contribution | Classifier Performance



Contribution | Classifier Performance



Contribution | Classifier Performance



Conclusions

We have implemented and tested the model proposed by Zafarani et Al

On a updated and different dataset, differing by

- Quantity
- Data sources
- Data homogeneity
- Predisposition of data to being classified

  Zafarani Baseline method scored 77% Accuracy

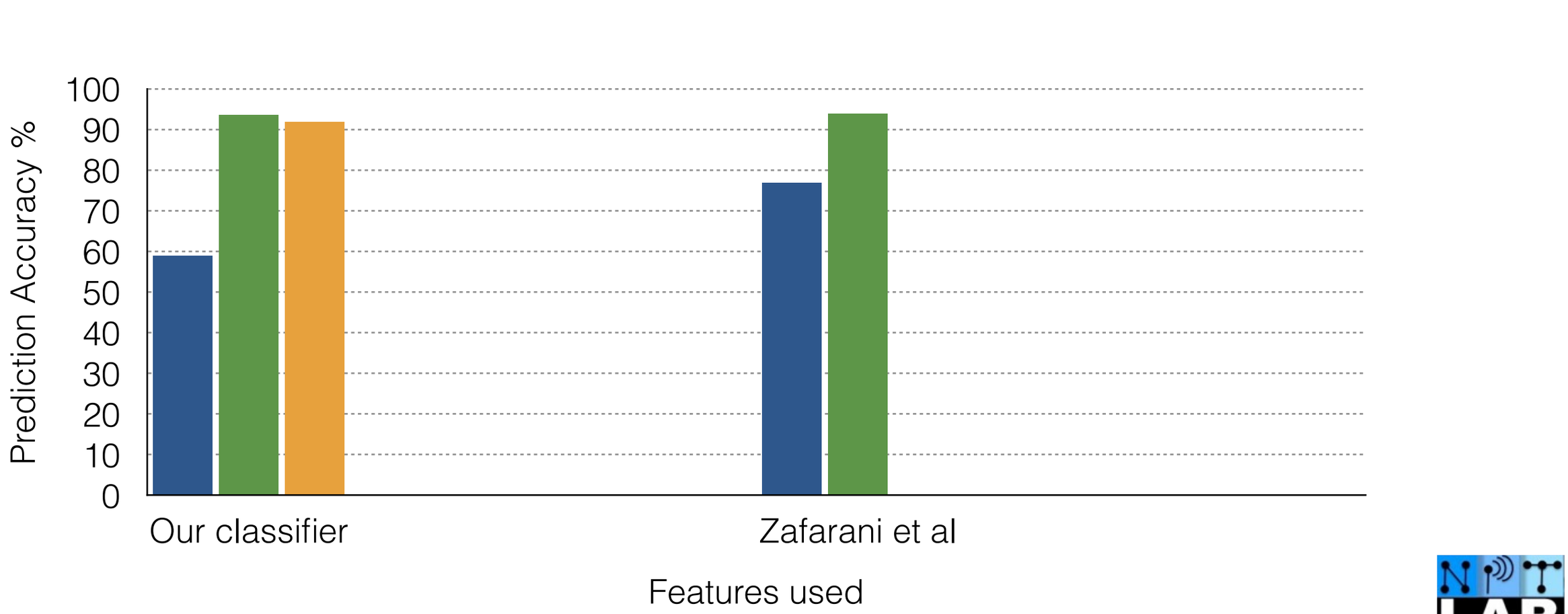
  This work 59%

Conclusions | Comparison with the state of the art - Zafarani et Al

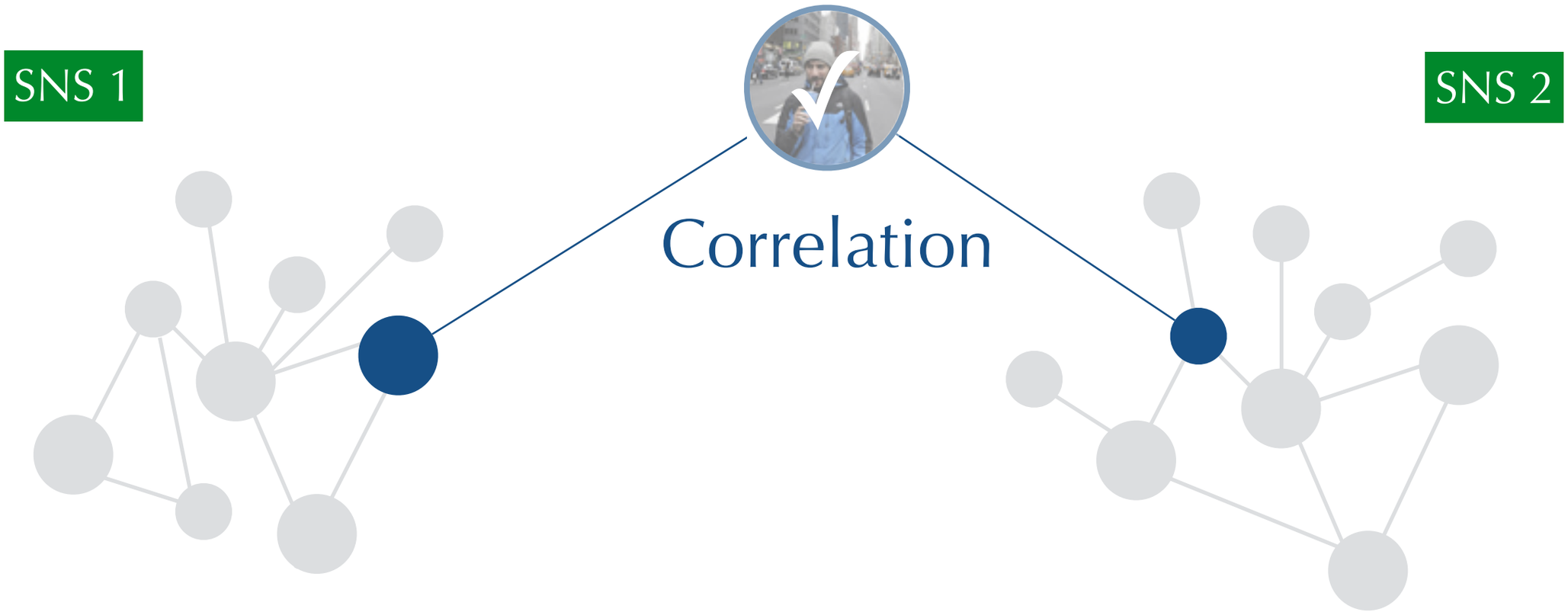
Proposing some alternative features

Measuring an accuracy performance of ~94%

in line with the results obtained by Zafarani et Al



Problem | Identify an individual through different social network services



Thanks | Any questions?

Thank you for your attention