

People across Social Media Sites: Usage and Behaviors

Matteo Zignani, Mattia Dimauro, Sabrina Gaito, Gian Paolo Rossi

Network Protocols and Technologies Lab
University of Milan



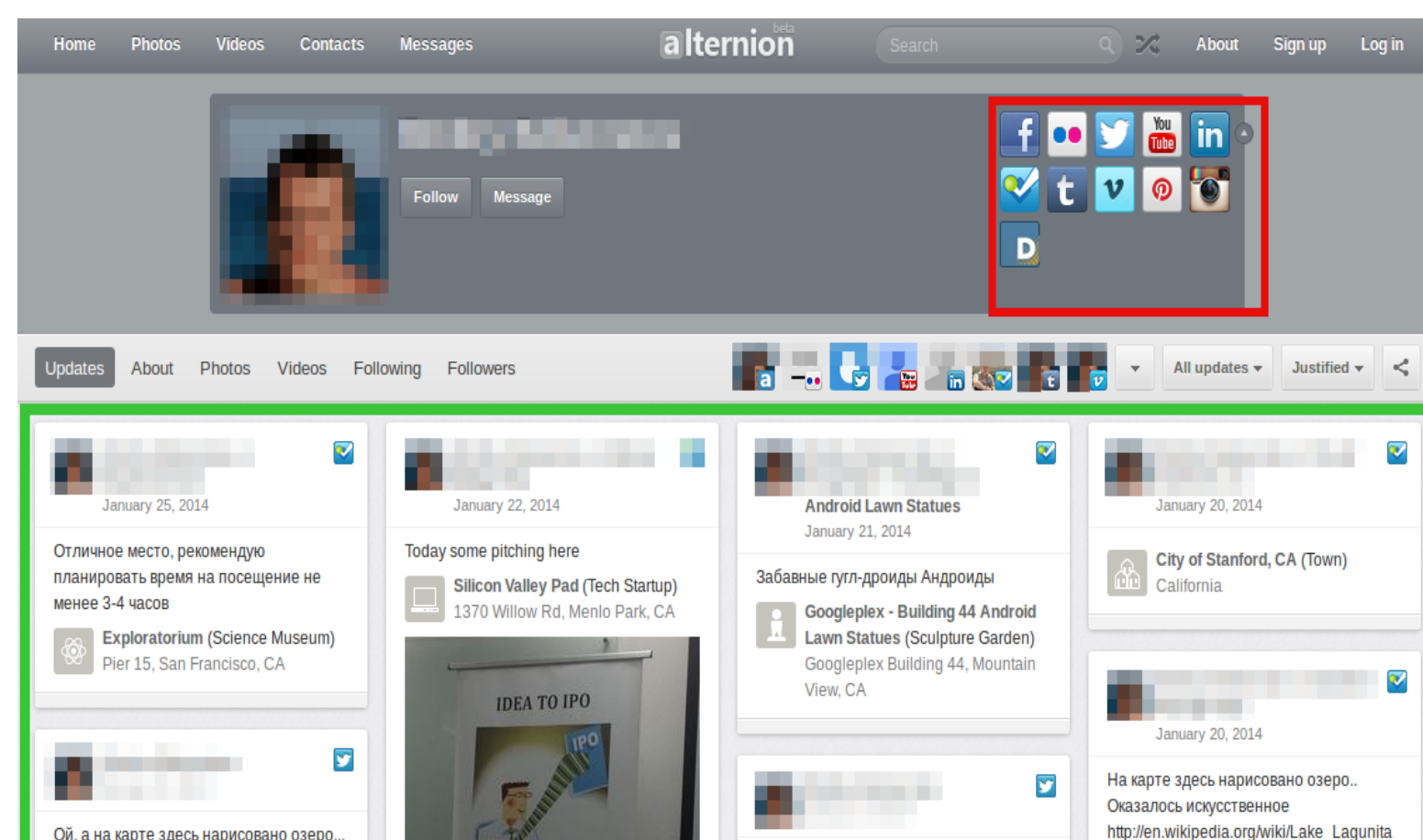
Motivations

It is becoming evident that individuals are used to express their sociality through **multiple layers**, each associated to a different medium. Users have access to a wide portfolio of social services which allow them to differently express their interests and convey diverse **contents**.

People have recently shown a growing interest in tools managing their online life in a centralized way, i.e. **social media aggregators**. Social media aggregators allow users to combine their own identities into a single profile.

Data collection

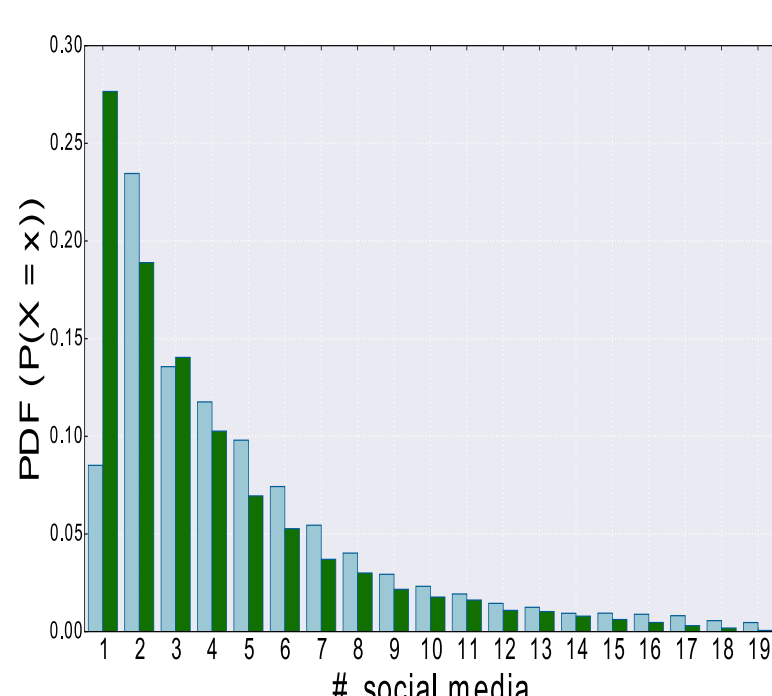
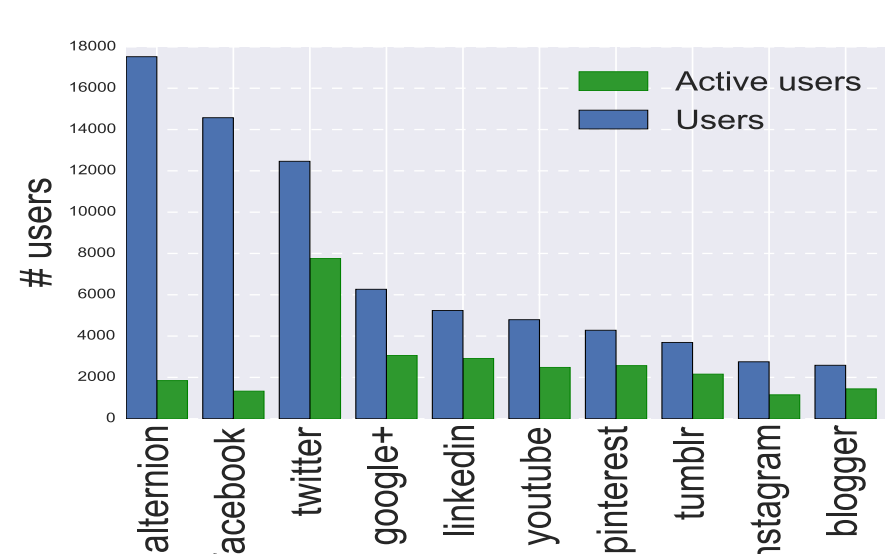
The aggregation is enabled by APIs provided by social networks => people have control on the data the aggregation platform can access.



Most of these services offer a private space to the users but public **profiles** and **content** are not available. The media aggregator **Alternion** allows users to decide which information make public.

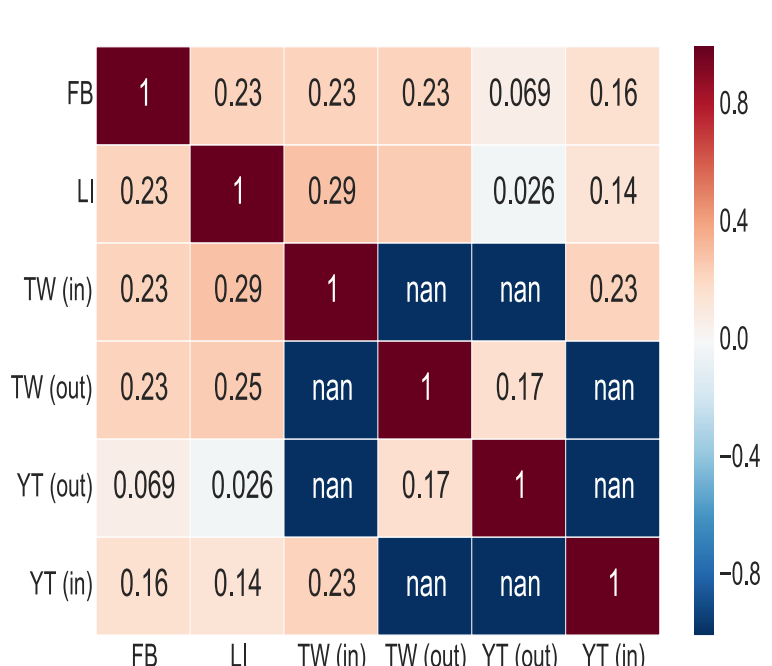
Social Media Usage

- Facebook is the most used social media and the top 10 corresponds to the most popular social media.
- About half of the profiles (9829) are active and do prefer Twitter as publishing media.
- Facebook active users are more productive than Google+ users



Q1: How common is the use of multiple social media sites? Does the same hold for active users?

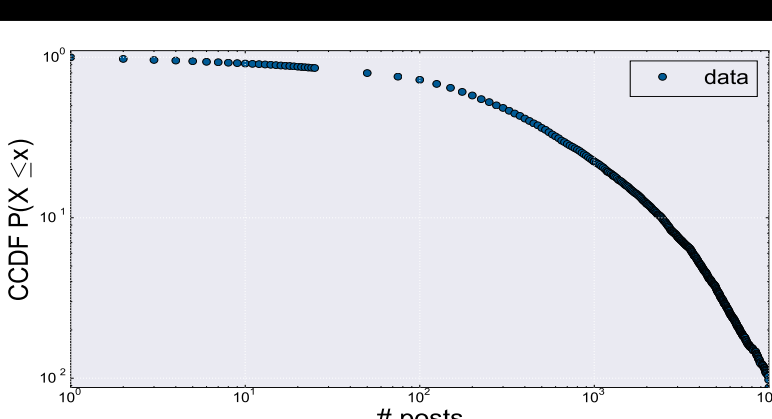
On average a user is **simultaneously** registered on **5** social media sites, while more than 95% of users have joined at most 17 platforms. 73% of active users publish on **at least two** social networks.



Q2: Is a person's popularity more or less equally distributed across the social media ?

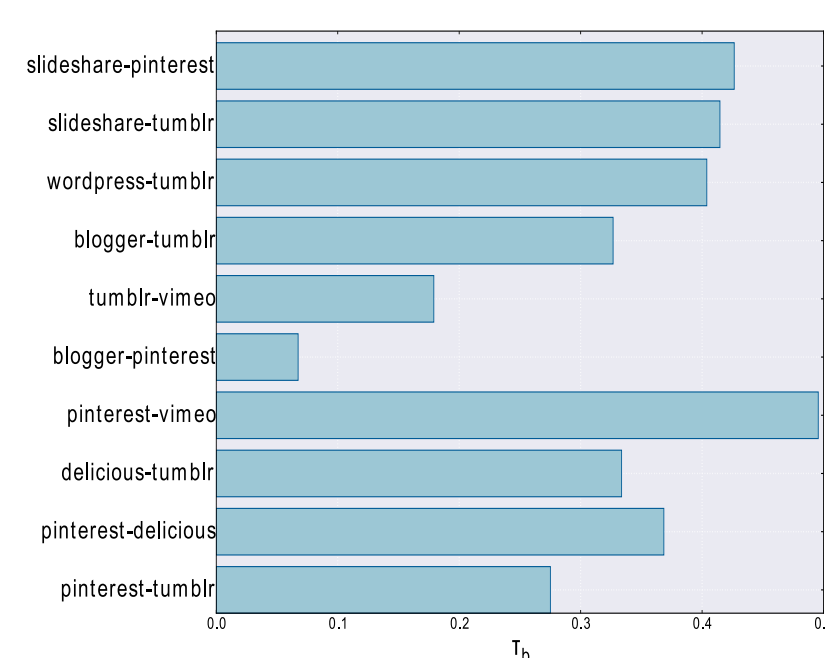
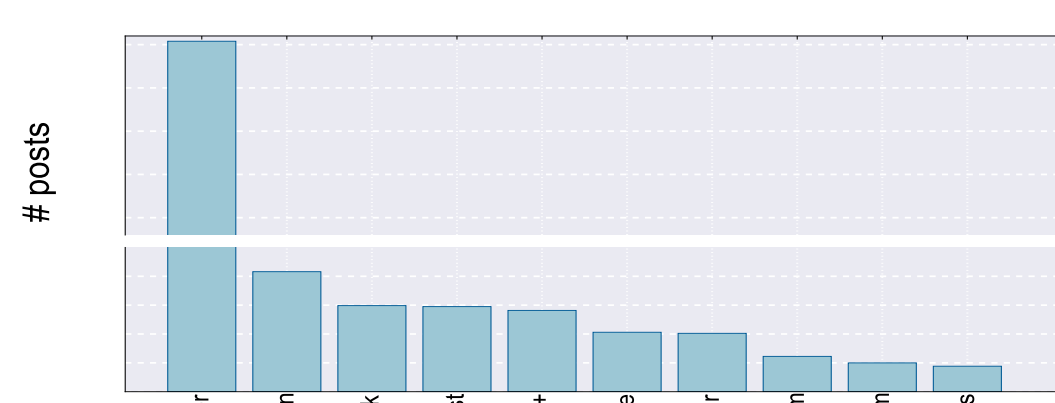
Users may have a very **different popularity** across the services: a single user might be an hub on one system and loose part of its hubbiness on the other.

Posting on Social Media



Number of contents published per user: Distribution: heavy tail with exponential cut-off. On average about 800 different contents and half of the users produced > 300 updates.

Predominant role of Twitter in the production of contents and posts, more than 6 million posts over 8.5 million events.



Q3: Is users' production alike in different media? Do most active users behave similarly?

Correlations between the posting rates on different social sites show that an easy answer is not possible, although we measure **slightly positive correlations** for many couples of social platforms.

Dataset Features

Dataset 1 (D1):

- Number of social media: 152; from Facebook, Twitter to Discus, Zazzle.
- Number of Alternion accounts: 19680.
- Number of content published by API: 8.5 millions post.
- Time interval from 10 July 2005 to 10 November 2104.

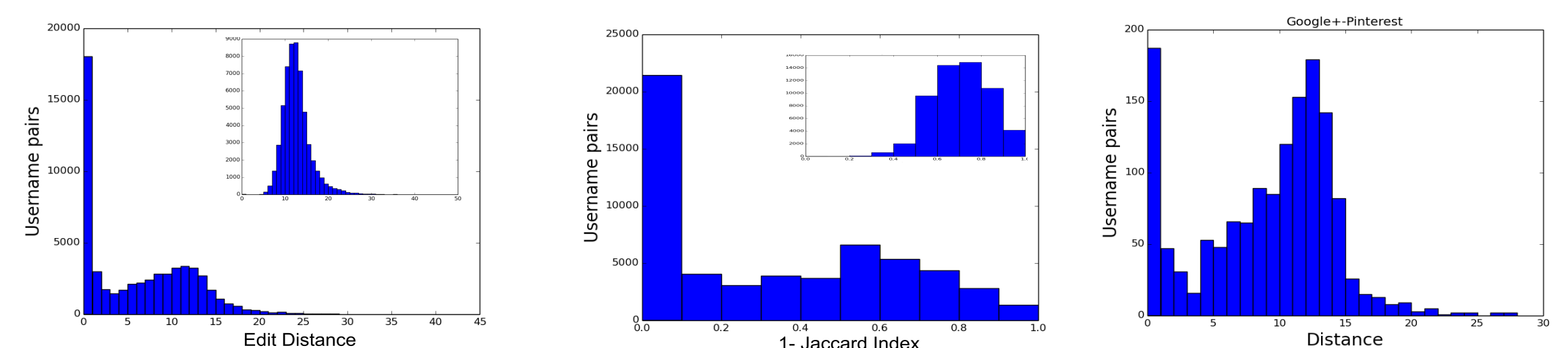
Dataset 2 (D2)

- Account with English first name: 15.000.
- Username, first and last name for each account associated to an individual.

Patterns in Username Usage

D1 and D2 sum up properties about the choice of the username the first way of being identified and an element in the self-presentation.

On D2 we evaluate how users are coherent in the choice of the username by computing the edit distance and the Jaccard coefficient on the username pairs associated an individual.



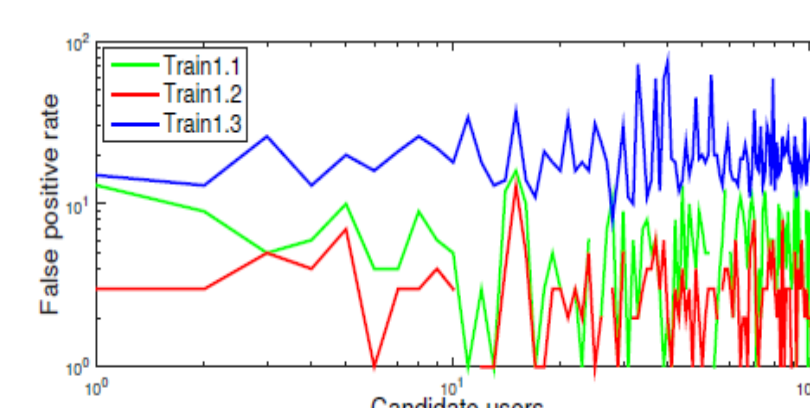
Peak at 0: People are not likely to change their username

This behavior depends on the pair of social media we consider => the change of the username between Pinterest and Google+ is more evident w.r.t. the average behavior.

Application: Profile Matching

We introduce an approach to find a **mapping** among **identities** across OSNs based on minimum common fields:

- Act on the training and test datasets to increase the homophily of negative instance.
- Reduce the number of false positives.



Train dataset	Test dataset	Random Forest				Multilayer Perceptron			
		Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
Train1.1	Test1.1	0.962	0.964	0.962	0.962	0.96	0.961	0.960	0.960
	Test1.2	0.606	0.614	0.607%	0.571	0.584	0.581	0.584	0.551
Train1.2	Test1.1	0.926	0.933	0.927	0.927	0.927	0.936	0.928	0.927
	Test1.2	0.796	0.840	0.797	0.794	0.765	0.787	0.766	0.765
Train1.3	Test1.1	0.890	0.890	0.890	0.870	0.916	0.926	0.916	0.916
	Test1.2	0.787	0.813	0.788	0.787	0.777	0.831	0.778	0.774
Train2.1	Test2.1	0.954	0.955	0.954	0.954	0.957	0.958	0.957	0.957
	Test2.2	0.847	0.848	0.847	0.845	0.844	0.850	0.845	0.842
Train2.2	Test2.1	0.933	0.935	0.933	0.933	0.919	0.921	0.920	0.919
	Test2.2	0.909	0.913	0.910	0.910	0.907	0.914	0.908	0.908
Train2.3	Test2.1	0.897	0.898	0.898	0.898	0.901	0.901	0.901	0.901
	Test2.2	0.917	0.925	0.917	0.917	0.911	0.913	0.911	0.911

We evaluate the **overlapping** of the **neighborhoods** of the same person in Google+ and Facebook. The average overlapping is **low** => users exploit different social media to establish relationships with different individuals.