Anatomy Of the TWITTER Social Graph

Source

"Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph"

An **ACM-2014** Paper by **Maksym Gabielkov**, **Ashwin Rao** and **Arnaud Legout** from Inria, France.

Flow

- Introduction
- Crawling Twitter
- Macro Structure: Creation
- Macro Structure: Analysis
- Time Evolution

Introduction

Introduction

- Unlike Facebook, the Twitter graph is **directed**.
- Twitter data of **505 million accounts** and **23 billion edges** was collected by the authors.

.

  - The accounts are nodes and an edge **A->B indicates** that user A **follows** user B on twitter.
  - The information **flow** happens along the **opposite** direction of an edge.

.

Basic Terms

- **Followers -** The users following the subject node.
- **Followings -** The users followed by the subject node.
- **Protected account -** The accounts which require approval to access the followings list.

Crawling Twitter

Crawling Methodology

- The crawl was done using REST API 1.0 from **March to July 2012**.
- The crawl was done
  - Distributed crawling on **550 normal** machines.
  - **2** whitelisted machines with **20K, 100K requests** per hour each.

Crawling Methodology

- ➔ Twitter assigns **IDs** with **non-contiguous** numbers.
- ➔ The crawl was based on user IDs. An upper bound to the account ID was found.
- ➔ Requests are sent with each possible ID till the upper bound and for a valid ID, we extract the **list of followings** for **non-protected** accounts.

Crawling Limitations

- The API restricts **normal machines** to **150** requests **per hour**.
- About **6%** of the accounts were **protected** and were not a part of the crawl.
- The accounts **deactivated/suspended** during the crawl were **not** in the dataset.
- All in all 6.3% of twitter is not present in the graph.

Validating Crawling

To validate the graph constructed from the followings list of nodes

- ➔ The **number of followers and followings** were got from the **API** again **in** a **recrawl**.

- ➔ The difference in followers and followings, through both methods was computed for each node and plotted.

Comparison graph

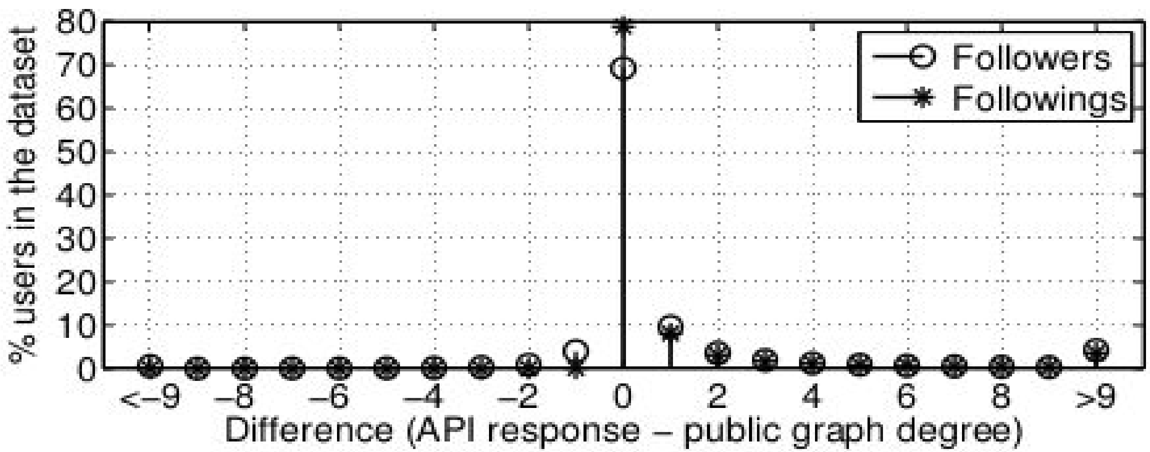


Figure 1: The difference in number of followers and followings between the data from user accounts and the public social graph reconstructed from our dataset.

Crawling Comparison

- **Most** of the accounts have **same** in **both** the **crawls**.
- But there are a few discrepancy because
  - Protected accounts are discarded in the initial construction.
  - The **time** of crawls **differ**.

MacroStructure Creation

Approach

“

The **graph** is **too large** for analysis and hence has to be **reduced in size** to **understand** its **nature**.

Generating Macrostructure

- Firstly, we find all the **connected components** and **group** them **as** a **single node**.

- The **edges** are **replaced** with weighted edges equal to **number of edges** being **replaced**.

- The graph reduces to a **DAG** with **half the size of nodes**, still it is a big one to analyse.

Generating Macrostructure

- There is a single **largest** component LSC which has about **50%** of **nodes**.
- Run a **BFS** from **LSC** and group the nodes **encountered** in **OUT** component.
- Then run a **reverse BFS** and group the nodes **encountered** as **IN** component.

Generating Macrostructure

- ➔ BFS from IN -> IN TENDRILS
- ➔ Reverse BFS from OUT -> OUT TENDRILS
- ➔ Nodes encountered in both -> BRIDGES
- ➔ Remaining Nodes
  - ◆ If they have a **link** to **nodes** in **other components** they are put in **OTHER** component.
  - ◆ If they are **not connected** to any **earlier component**, it is put in **DISCONNECTED** component.

MacroStructure

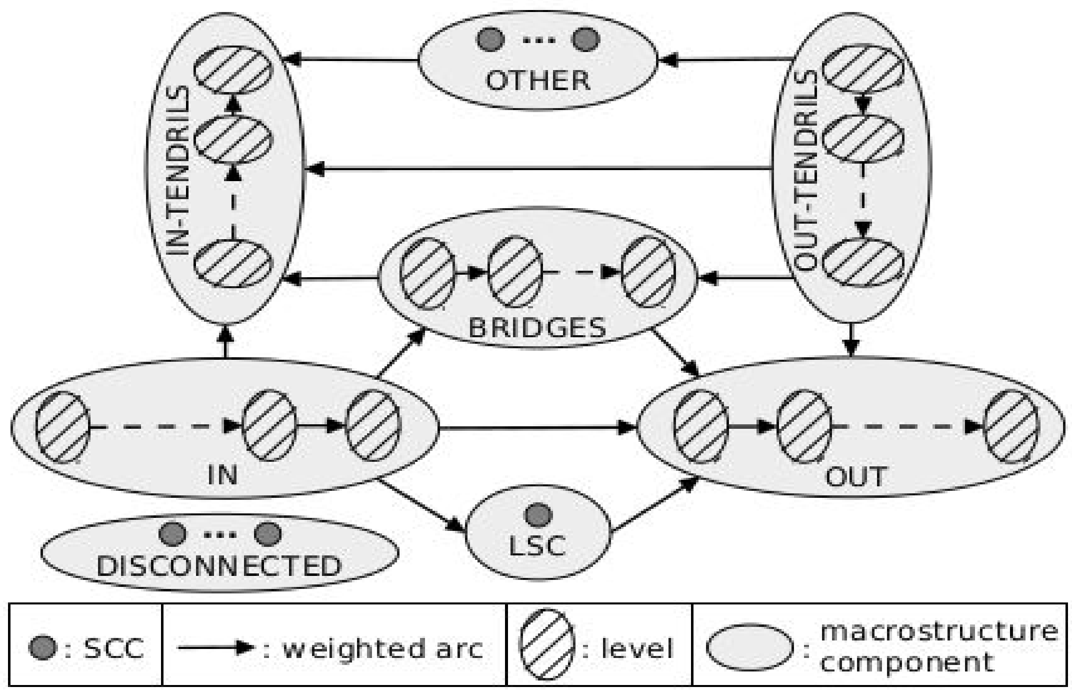


Figure 2: Macrostructure of any directed graph.

MacroStructure

Analysis

Terminology

➔ In each of the components of the macrostructure, we will be analysing about three types of accounts

- **Regular accounts -** The accounts with normal twitter activity.
- **Abandoned accounts -** The accounts with few followers and followings and no recent tweet activity.
- **Suspended accounts -** The accounts which have been terminated by Twitter for violation of terms of use/malicious use.

Account Distribution among Components

Component	Top followed (%)	Top following (%)	Top tweeting (%)	Experts (%)	Verified (%)
LSC	96.95	100	88.66	94.28	97.01
OUT	3.05	0	10.79	1.33	2.99
IN	0	0	0.07	0.01	0
DISC.	0	0	0.47	0.01	0
OUT-T.	0	0	0	0	0
IN-T.	0	0	0	0	0
BRID.	0	0	0	0	0
OTHER	0	0	0.01	0	0

Distributions Among The Components

	Arcs (%)		Tweets (%)	Accounts (%)
	followers	followings		
LSC	98.01	96.13	98.05	50.71
OUT	1.96	0.02	1.49	5.30
IN	0.02	3.83	0.25	21.36
DISC.	<0.01	<0.01	0.21	21.60
Others	<0.01	0.02	<0.01	1.03
Total	23×10^9		127×10^9	505×10^6

Table Data Analysis

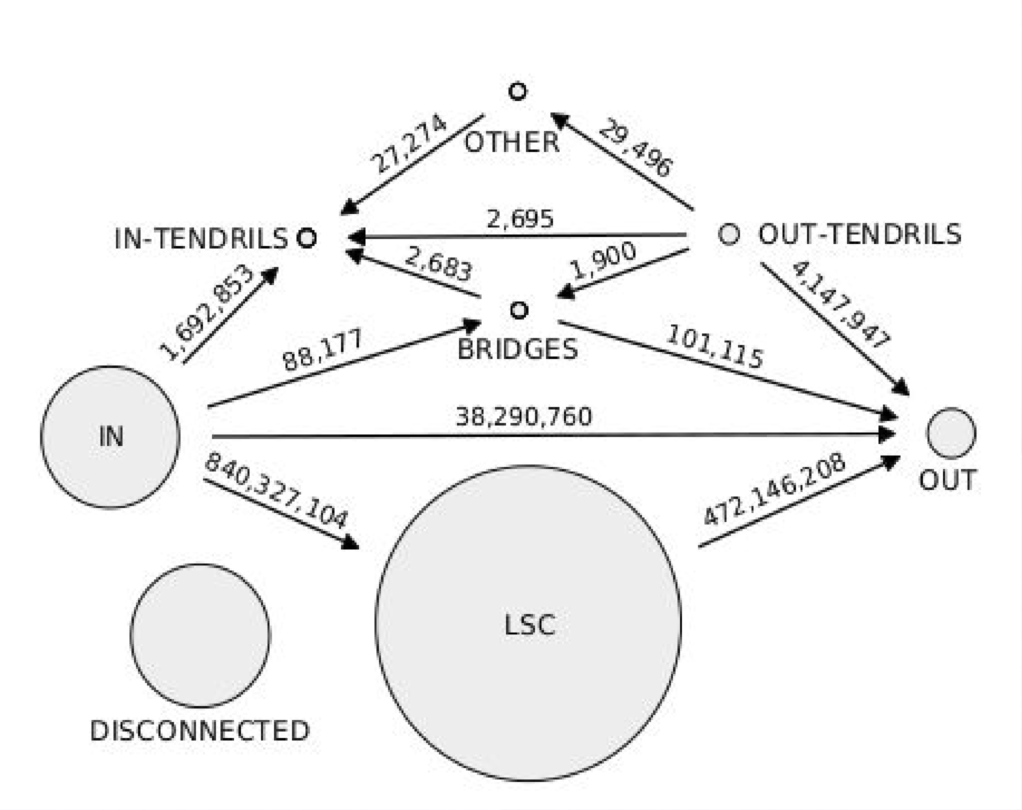
- LSC -> Most important component in terms of activity but has only 50% nodes.

- **LSC, IN, DISCONNECTED and OUT** from about **99%** of the graph.

- The general trend is
  - IN - has no/less followers
  - OUT - has no/less followings
  - DISCONNECTED - has no/less both

LSC

Regular users



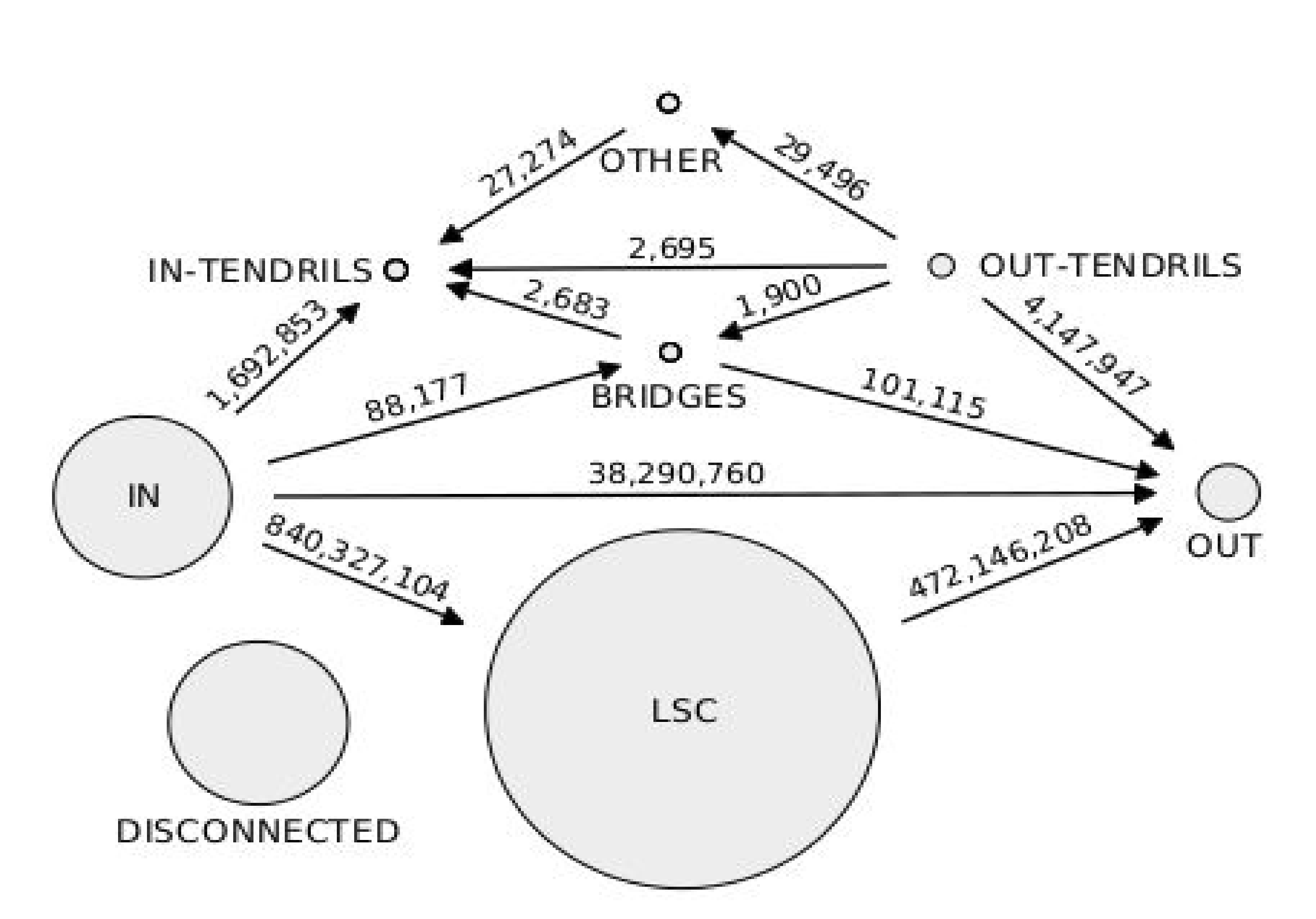
The Graph with component Sizes and edge weights

LSC

- **Most** of the accounts are **Regular.**
- The accounts with very **few followers and followings** and no recent tweets are **Abandoned**.
- 98% of the accounts with **high followings and <=2 followers** are **Suspended**.
- **High** Tweet **activity** and **<=2 followers** ->**Bots** tweeting and making **interface to third party data.**

OUT Component

Selfish Users



The Graph with component Sizes and edge weights

OUT Component

- It has **6%** of the total nodes.
- The accounts in this component **do not follow** accounts **outside OUT component.**
- **Most** of the regular accounts in here are **Selfish** i.e they **do not follow other accounts** and rather use it as a publish media.
- This is a **dying trend** and over the years the **share of OUT** has been **decreasing.**

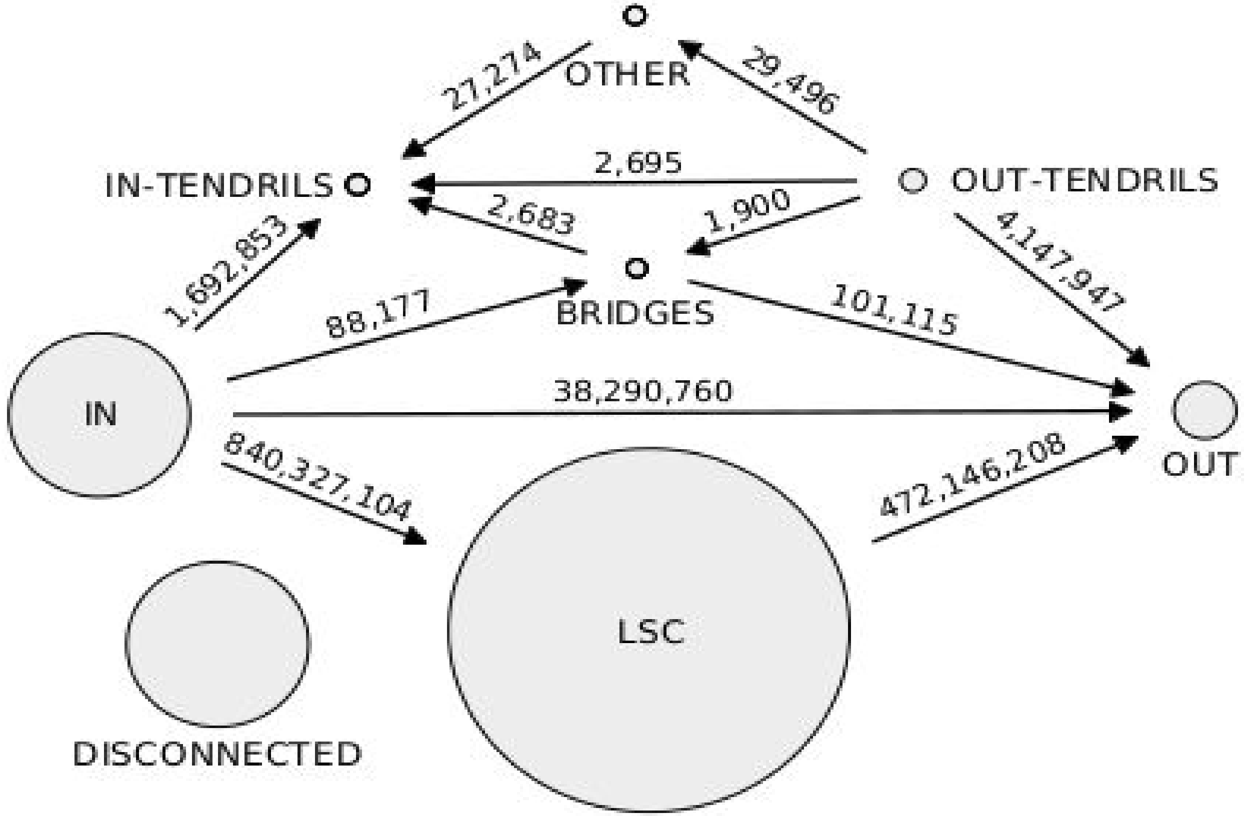
OUT Component

- ➔ **Abandoned** accounts **in OUT** are mostly the account from the DISCONNECTED component and are **moved** here by a **single/few followers**.

- ➔ **Smallest no. of malicious accounts** as the malicious accounts generally spam by **increasing followers** which is **not** the case **here**.

IN Component

Passive USers



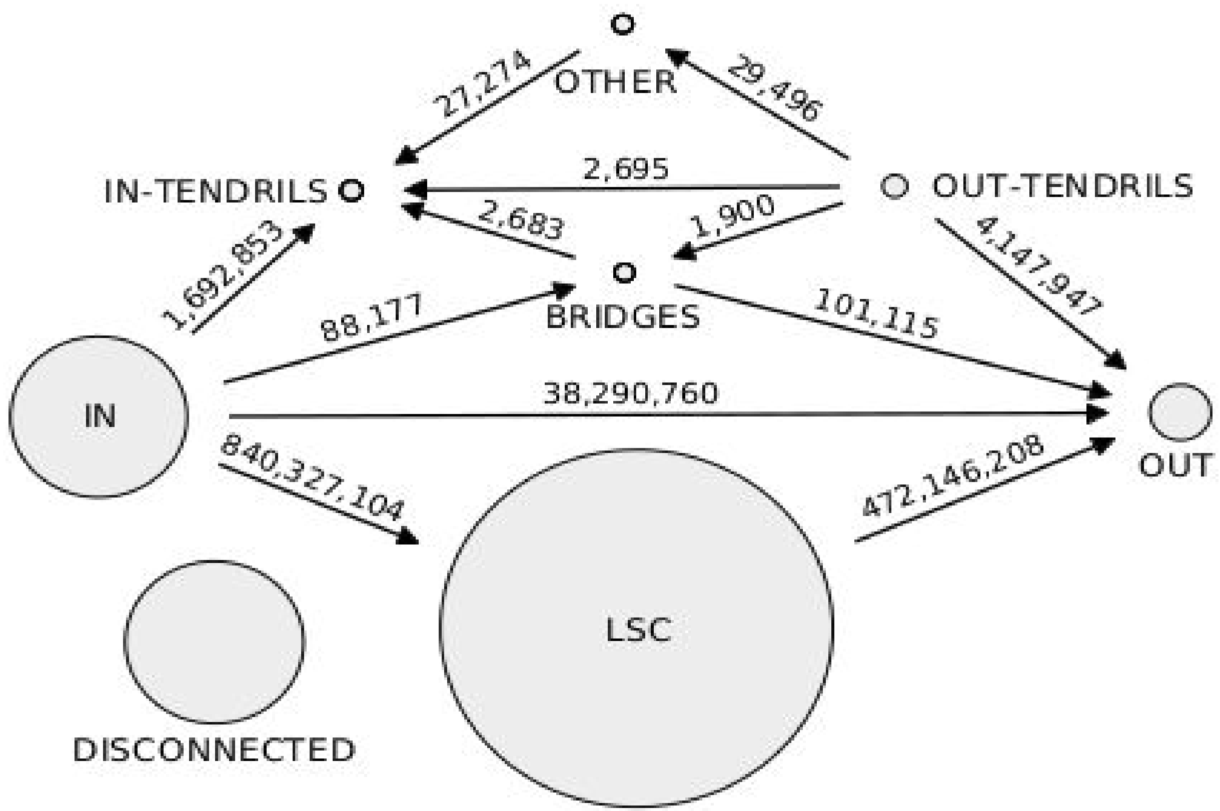
The Graph with component Sizes and edge weights

IN Component

- ➔ It has **20%** of the total accounts.
- ➔ **Large** amount of **Suspended** accounts as the nodes have a lot of followers and very few/no followings.
- ➔ The **Regular** accounts in IN are **passive users** who use it as **information media**.
- ➔ The **accounts** are **young** indicating that most of them **migrate** in **future**.

DISCONNECTED Component

Abandoned users



The Graph with component Sizes and edge weights

DISCONNECTED Component

- ➔ Most of the accounts here are **Abandoned** accounts.
- ➔ It has **20%** of the nodes and hosts a large fraction of malicious activity on Twitter.
- ➔ It is also a **transit** place for **very young accounts**.
- ➔ Most accounts have no followings, no followers, and no tweets.

Other 4 components

1% nodes only

Transit to LSC, IN, OUT

Some things to ponder

Macrostructure

The macrostructure constrains the propagation of information.

Reasonably simple model

Work sheds light on how to abstract the social graph, with only 3 main components with active accounts.

Influencers identification

Identification of roles helps us to identify real influencers by exclusively focusing on followers in the IN component, as no. of followers cannot be a real metric as users perform link farming to inc. no. of followers.

Correlation and data

B/w components in macrostructure and usage of accounts in these components.

Thus, identifies spots for applying graph techniques.

Eg. All sampling techniques following arcs from active accounts will miss malicious activity in DISCONNECTED component.

Time Evolution

Method Of Generating At Different Times

- ➔ We have the crawled dataset in 2012.
- ➔ Now to discuss evolution from Jan 2007-July 2012.
- ➔ To get the Social graph at any time D
  - ◆ **Remove** all the **nodes** created **after** time **D**.
  - ◆ **Remove** all the **edges related** to the **removed nodes**.
  - ◆ Apply the macrostructure extraction discussed above.

Limitations of the approximation

- Suspended and deactivated accounts info is not accounted for.
- The edge creation(follow link) can happen at various times but we are considering that A->B is formed at the creation time of the youngest account.

Validating with other old data sets

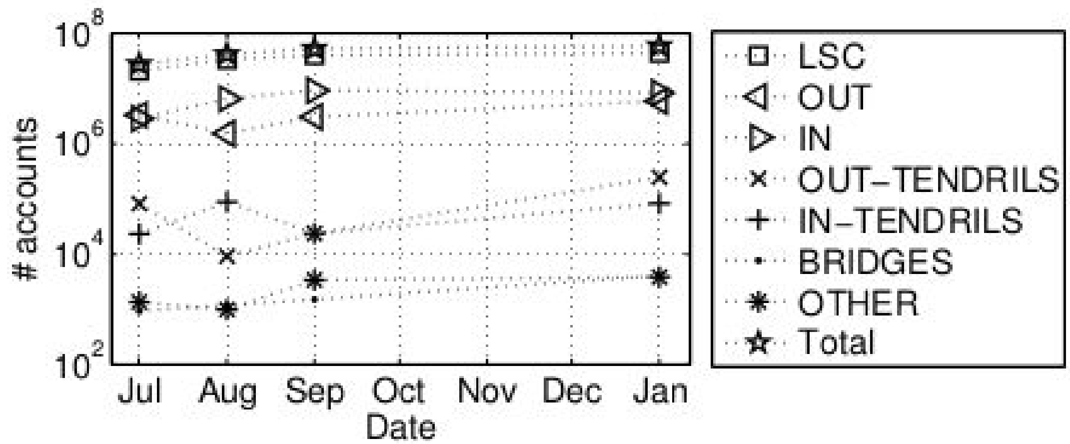
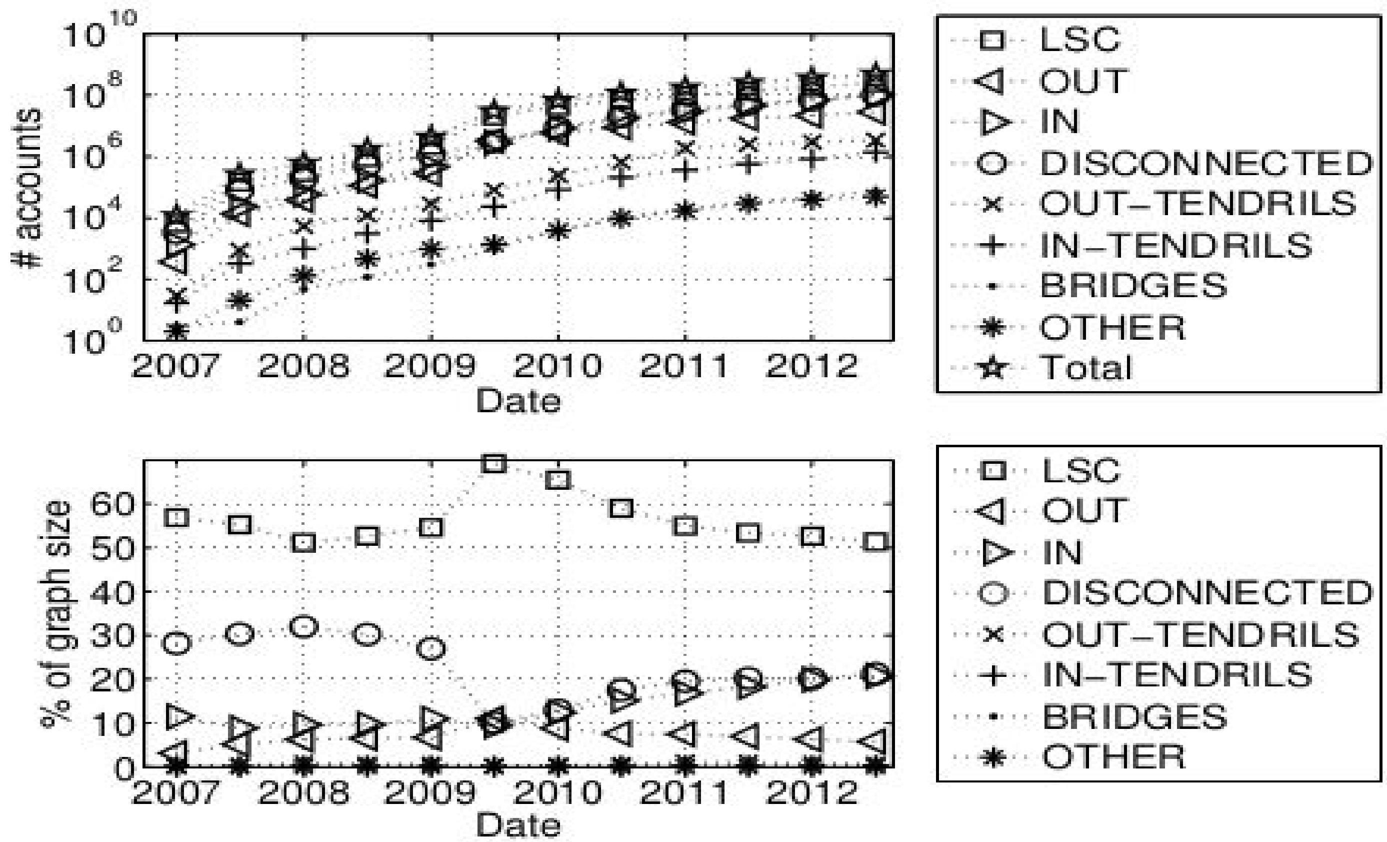


Figure 8: Comparison of our estimated graphs of 2009 (labeled Jul and Jan) with two existing Twitter datasets made in August [16] and September [9] 2009. *Our simple methodology gives an approximation of the macrostructure of the Twitter social graph that is consistent with existing datasets.*

Analysis

- Twitter Social graph has been constructed for every six months from January 2007 to July 2012.
- The macro structure was extracted and the evolution of sizes of various components was analysed.

Evolution of the Components



Analysis

- **2009** marks a **significant change** in the Twitter usage.
- Numbers of users boomed from **4.2 million to 67.8 million** in 2009-2010.
- We see that LSC has shooted up and DISCONNECTED component share decreased indicating a positive boom.

Analysis - Post 2009

- We also see that in the recent years the **OUT** component **share** is **coming down** which is indicative of **decrease** in the **Selfish user** trend.

- We also see that **IN** component **share** is **increasing** indicating **increase** in **passive** users.

Thank You!

Any questions?

Bakhtiyar Syed
(syed.b@research.iiit.ac.in)

Krishna Chaitanya Pappu
(krishna.pappu@students.iiit.ac.in)

End Of Presentation