# LASTA: Large Scale Topic Assignment on Multiple Social Networks

Nemanja Spasojevic, Jinyun Yan, Adithya Rao, Prantik Bhattacharyya
Klout, Inc.
{nemanja, jinyun, adithya, prantik}@klout.com

# Highlights and Contributions

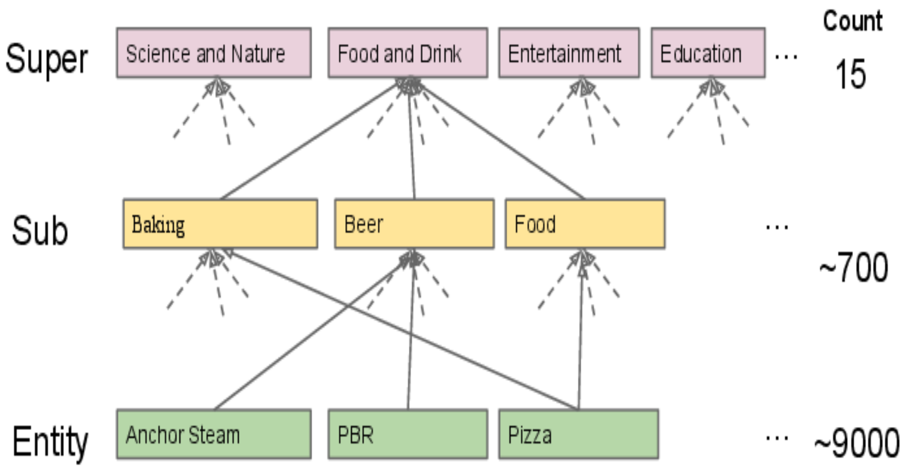
- **Fully deployed production system** to assign topics at scale
  - ~10,000 topics assigned to hundreds of millions of users daily
  - Reactive to fresh data
- Data from **multiple social networks** used to create an aggregated profile for a user:
  - Twitter, Facebook, LinkedIn, Google+, Wikipedia
  - User activity, profiles, connections
- **Feature engineering** approach that uses following categories:
  - Original generated content
  - Reactions to original content
  - Indirect attributions to user
  - Graph based features
- **Cross-Network information** leads to:
  - More topics assigned per user
  - More users who can be assigned topics
  - Better user-topic associations compared to using a single network

# Klout

- Klout is a social influence measurement tool.
- Users register on Klout.com and connect their social network accounts.
- Klout collects authorized/public information from connected networks.
- Klout derives influence scores and topics for users from collected data.

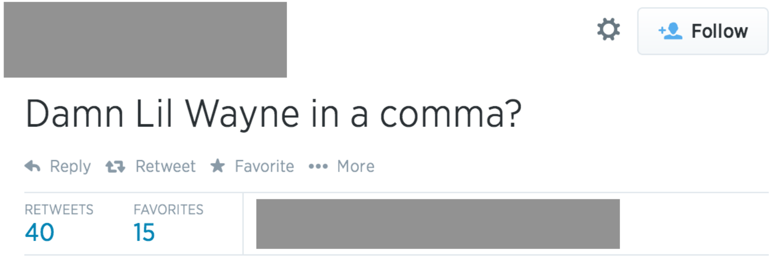
# Motivation

- Assign topics to the **long tail**
- Focus on **socially recognizable topics of interest**
    - Warren Buffett may be interested in *Ukulele* and *Online Bridge*, but is known for his recognizable interests like *Business* and *Money*.
- Applications **in Recommendation** and **Targeting** systems:
    - Content recommendations
    - User targeting
    - Question Answering
- **Extensibility** in terms of data sources.



# Challenges in social data

- **Message size:**
  - Overall data size may be huge, but message size per user may be small.
- **Text Sparsity**:
  - Many users may be passive consumers of content.
- **Noise**:
  - Social content abounds in colloquial language, slang, grammatical errors, abbreviations.
- **Context**:
  - Need to expand context to get more information

# Why use data from multiple networks?

- Phrase usage on different social networks is different
- Phrase overlap across social networks is small
- Combination of networks provides more quantity and diversity of phrases used.

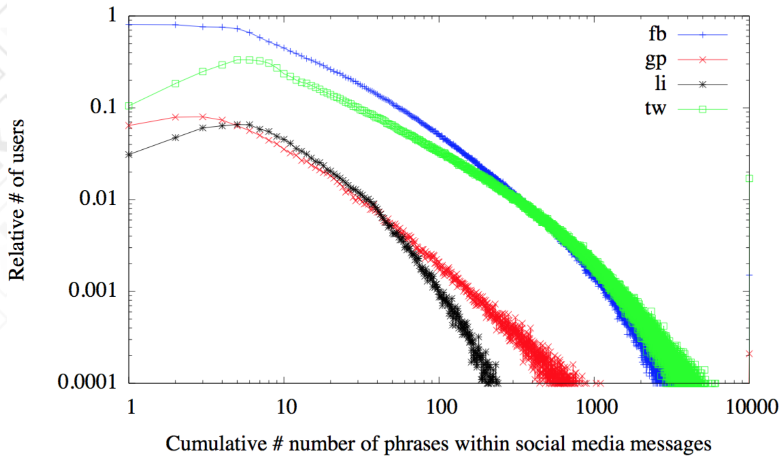



Fig. 1. Verbosity distribution across social networks

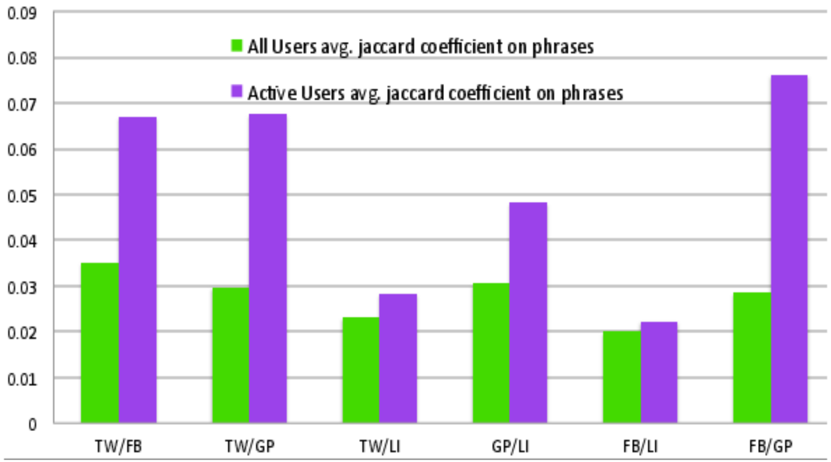
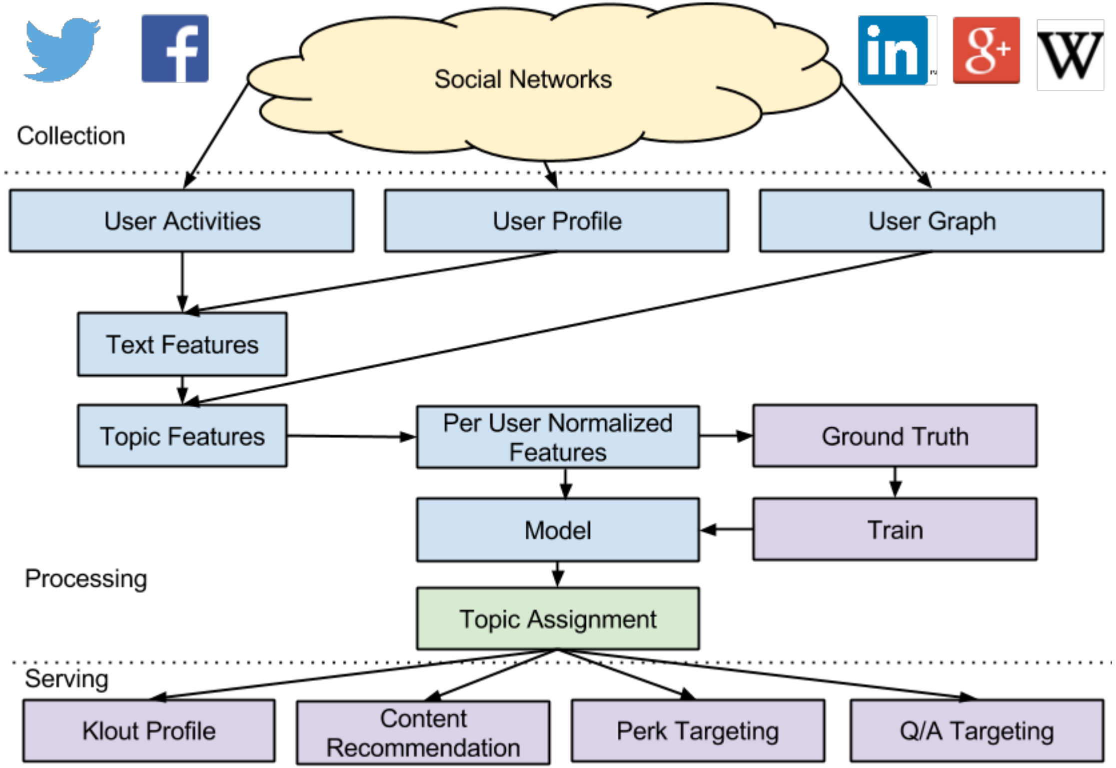



Fig. 2. Phrase overlap on social networks

# Data Pipeline

- **Facebook**: Authored status updates, shared URLs, commented and liked posts, text and tags associated with videos and pictures.
- **Twitter**: Authored tweets, retweets, mentions and replies on other tweets, shared URLs, created and joined lists.
- **LinkedIn**: Authored posts, comments, skills stated by the user and endorsed by connections.
- **Google+**: Authored messages, re-shares, comments, shared URLs and plus-ones.
- **Wikipedia**: Wikipedia pages for well known personalities.



# System Details

- Topic Assignment runs as a bulk job on the Hadoop MapReduce stack
  - HDFS, Hive, HBase

- Exploded Resource footprint (uncompressed reads/writes from HDFS):
  - Feature Generation: 55.42 CPU days, 6.66 PB reads, 2.33 PB writes
  - Score generation: 11.33 CPU days, 3.78 PB reads, 1.09 PB writes

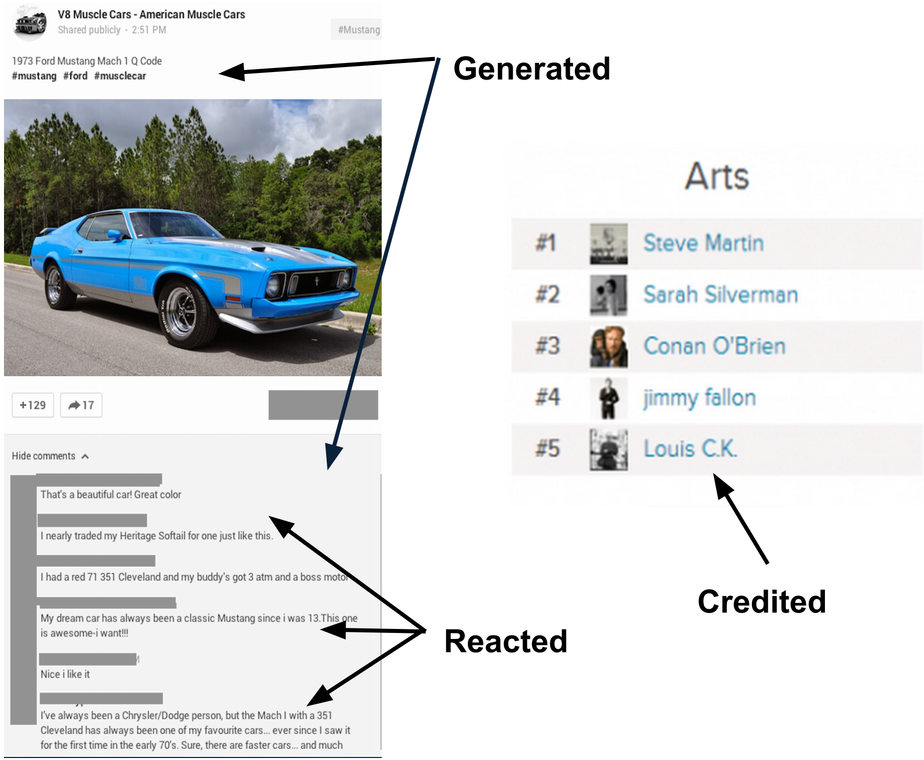
- Hive User Defined Functions (UDFs) implement utilities for data aggregation and transformation
  - https://github.com/klout/brickhouse

- Machine Learned models are trained offline and improved regularly.

# Feature Engineering

- A **Topic Feature** in the pipeline is represented as a bag-of-topics derived in a specific manner.
  - eg. TW_MSG_TEXT => { (topic1, 1.0), (topic2, 3.0), … (topicN, 1.0) }
  - A particular topic may occur in multiple bags of topics.

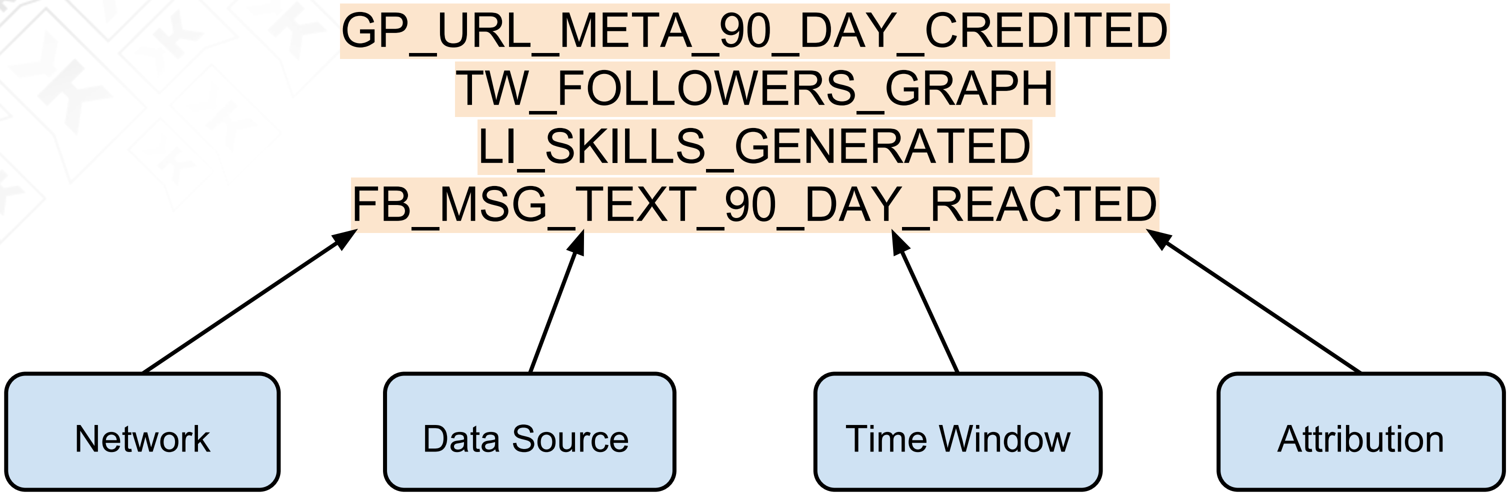
- Data sources are attributed to users as:
  - **Generated**: Original text, urls created and shared by a user.
  - **Reacted**: Reactions to original content from a user.
  - **Credited**: Attributions that do not depend directly on the activity of a user.
  - **Graph**: Topics derived for friends, followers, connections.



# Feature Engineering

- Each Feature is encoded as **<network>_<data-source>_<time-window>_<attribution>**

- Extensibility to create new features is important for experimentation and prototyping
    - eg. Add a new time window, or a new data source



# Ground Truth

- Since we want socially recognizable topics, members in a user's social graph evaluate topics for the user.
- Order is not considered during labeling.

| Statistics | Value |
|---|---|
| # of participants | 43 |
| # of evaluated users | 766 |
| # of (user, topic) labels | 32,264 |
| # of positive (user, topic) labels | 17,208 |
| # of negative (user, topic) labels | 15,056 |

Nemanja Spasojevic aka. sofronije

| _TOPIC_ | |
|---|---|
| water-polo | o |
| open-water-swimming | o |
| management | o |
| quantum-mechanics | o |
| c++ | o |
| klout | o |
| algorithms | o |

# Evaluation and results

Training:

- Transform bag of topics to a feature vector for each topic user pair (ti, u).
- Train a Binary Classification model using ground truth data.

Evaluation on test set:

- Single Network comparison
- Attribution Comparison
- Most Predictive Features

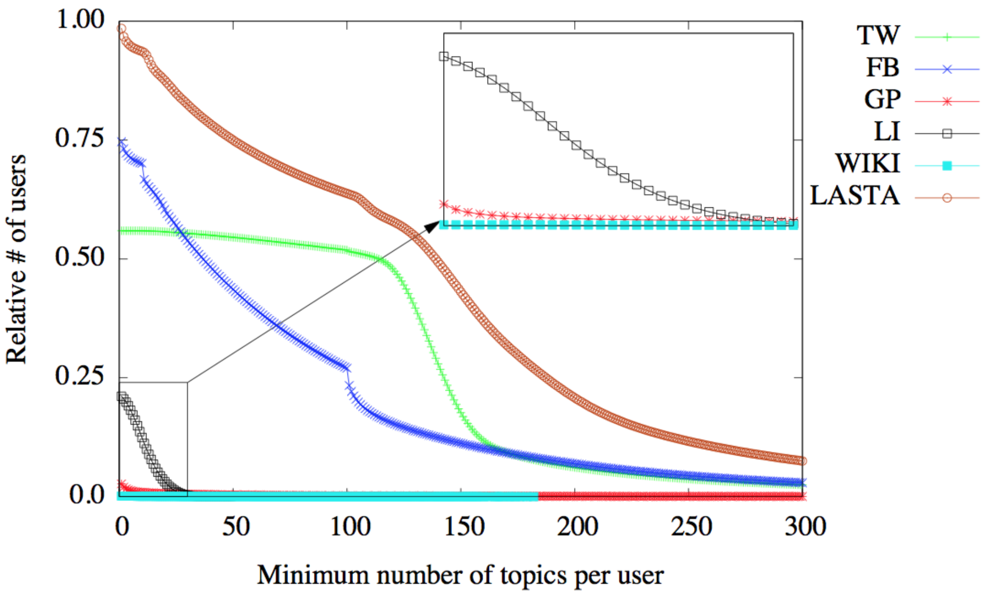
| User | Top 10 Topics |
|---|---|
| Marissa Mayer | yahoo, google, technology, business, twitter, social-media, flickr, design, marketing, seo, gmail |
| Lady Gaga | music, lady-gaga, celebrities, art, fashion, born-this-way, venus, entertainment, radio |
| Barack Obama | politics, affordable-care-act, health-care, new-york-times, congress, chicago, twitter, washington, illinois |

# LASTA vs single networks

**Better long tail performance:**
LASTA assigns topics to a higher percentage of users, compared to using a single network.

**More comprehensive per user:**

LASTA assigns more topics per user than a single network.



# Cross-Network topics

Table 9: Super-topic percentage distribution across different networks

| Super-topic | LASTA | TW | FB | LI | GP | WIKI |
|---|---|---|---|---|---|---|
| technology | 23.972 | 19.706 | 11.559 | 33.420 | 22.822 | 8.247 |
| entertainment | 23.987 | 20.049 | 20.866 | 3.406 | 14.377 | 30.669 |
| business | 15.893 | 10.628 | 7.567 | 41.053 | 12.857 | 10.937 |
| lifestyle | 7.910 | 7.403 | 11.409 | 2.328 | 7.969 | 4.810 |
| science-and-nature | 4.431 | 3.705 | 3.604 | 1.266 | 4.682 | 3.208 |
| arts-and-humanities | 6.605 | 7.056 | 6.836 | 5.765 | 9.392 | 13.373 |
| government-and-politics | 3.547 | 4.763 | 4.388 | 2.182 | 3.534 | 5.261 |
| sports-and-recreation | 4.379 | 7.503 | 7.591 | 0.659 | 4.913 | 7.921 |
| food-and-drink | 2.671 | 7.228 | 11.863 | 0.819 | 7.255 | 2.142 |
| health-and-wellness | 1.976 | 3.894 | 5.150 | 1.691 | 4.083 | 1.867 |
| fashion | 1.439 | 2.645 | 2.945 | 0.732 | 2.776 | 2.203 |
| education | 1.443 | 2.375 | 3.485 | 3.369 | 2.170 | 4.058 |
| news-and-media | 0.966 | 1.722 | 0.899 | 2.597 | 1.060 | 4.366 |
| travel-and-tourism | 0.535 | 0.779 | 1.155 | 0.614 | 1.041 | 0.654 |
| hobbies | 0.246 | 0.543 | 0.683 | 0.100 | 1.070 | 0.285 |

# Key Takeaways

- Do not ignore the long tail.
- Using more than one social network offers the opportunity to get a deeper understanding of users.
- Expanding context is important for topic derivation.
- If you are designing a production system, ensure it has the following characteristics:
    - It is extensible
    - It allows fast experimentation and prototyping

# Questions?