Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors

Conference of Online Social Network (COSN) 2015
2015. 11. 03.

Daejin Choi (djchoi@mmlab.snu.ac.kr)*
Jinyoung Han§, Taejoong Chung*,
Yong-Yeol Ahn¥, Byung-Gon Chun*, Ted "Taekyoung" Kwon*
Seoul National University*, University of California-Davis§, Indiana University¥

Reddit

- A platform supporting online communities for plenty of topics
  - **Every user** can make communities for **any topics** he/she wants
  - In a community (called *subreddit*), Reddit users communicate with others by posting / commenting in a **threaded way**

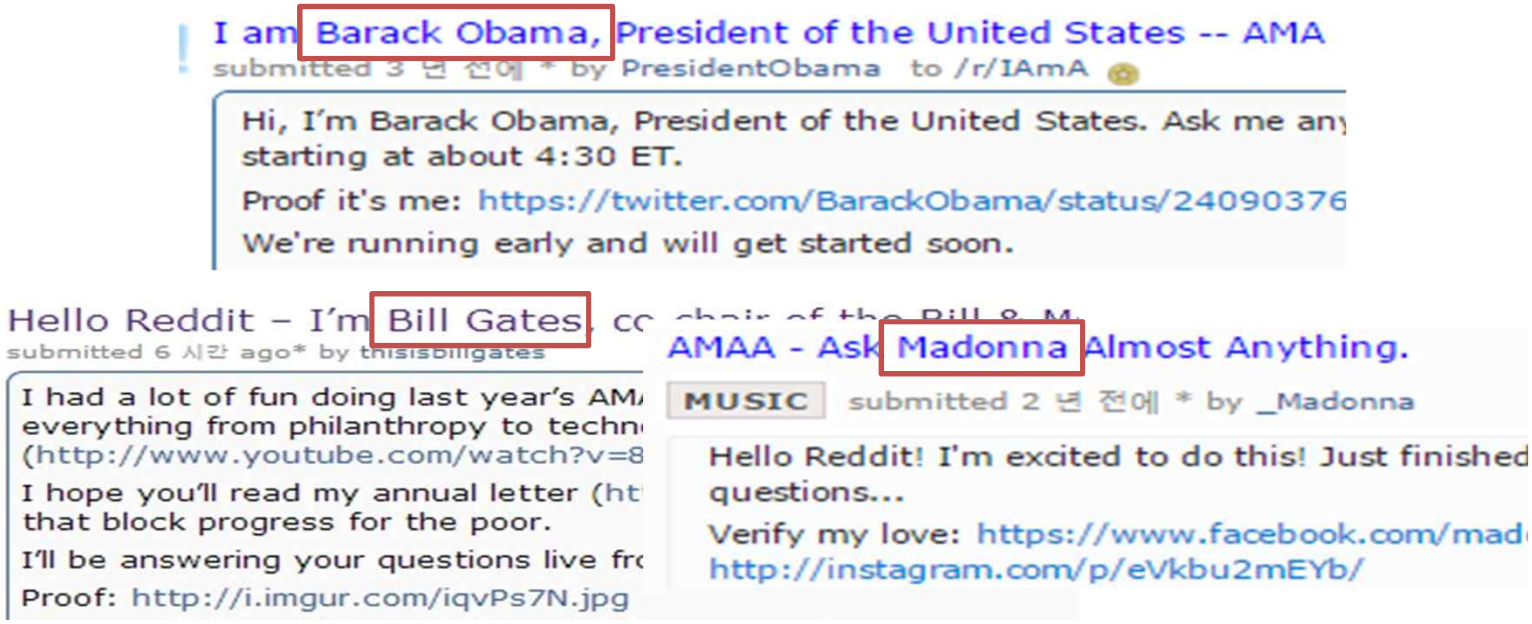
Post (External Link or Text)

Subreddit

Comments Hierarchy

Characteristics of Reddit

- **Popular** & **influential**
  - More than 169M unique users from 209 countries visit more than 7.5B pages in Reddit[1)]
  - 25th (World) and 11th (USA) most popular website[2)]

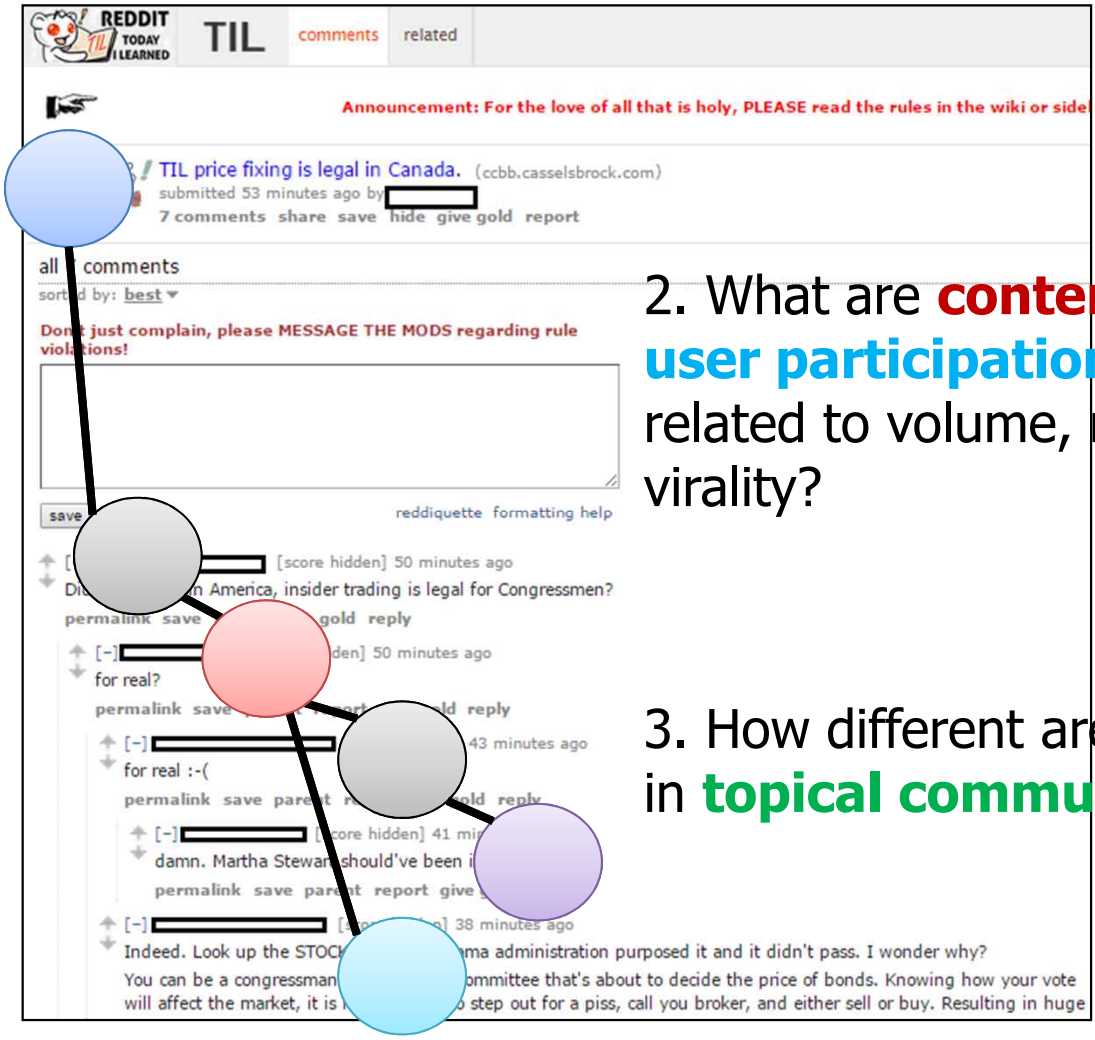


- Variety of topical communities & contents
  - More than 300K subreddits

1) reddit.com/about, May. 2015.
2) Alexa.com

Research Questions

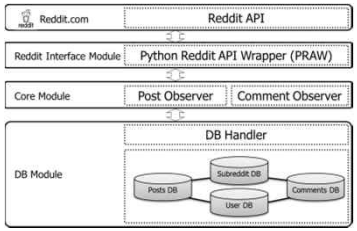
1. How can we characterize online (threaded) conversations?



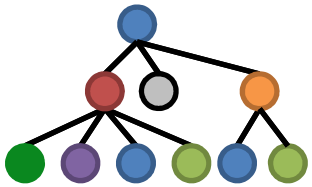
2. What are **content properties** & **user participation behaviors** related to volume, responsivenss, or virality?

3. How different are the characteristics in **topical communities**?

Measurement Methodology



Data Collection



Comment Tree Model

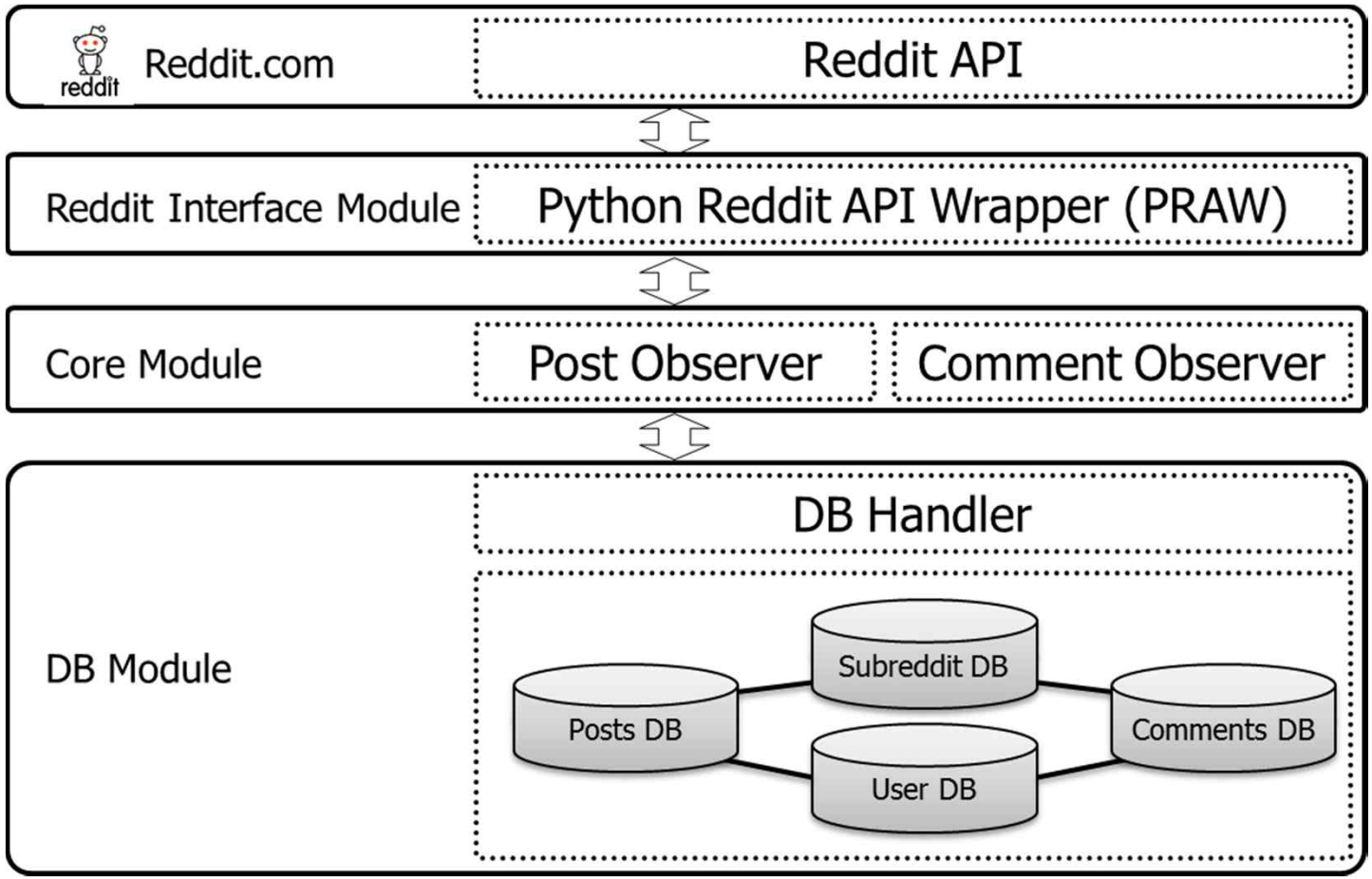
Comment Tree Analysis

Content Perspective

User Behavior Perspective

Community Analysis

Data Collection & Analysis System



- Crawling newly-uploaded posts/comments in top 100 subreddits by the number of subscribers
  - More than 60% of subscribers are covered

Period	2014-03-15 ~ 2014-04-18 (35 days)
Number of Users	1,531,247
Number of Posts	1,016,342
Number of Comments	18,626,530

Dataset Is Publicly Available!

To monitor the posts and their follow-up comments, we developed two key submodules in the core module: the post observer and comment observer. Once in every minute, the post observer monitors and fetches all new posts in each subreddit. At the time of our data collection, Reddit APIs provided up to 1,000 recent posts in each subreddit in the chronological order; hence our crawler fetches up to 1,000 posts every minute not to miss newly-uploaded posts. Whenever the post observer identifies a new post, the comment observer begins to keep track of all the comments relevant to the post. Similarly, the comment observer monitors and collect every comment associated with the posts that we have fetched. We collected every single post and comment during our measurement period since the observed maximum number of messages per minute was 722, which did not exceed the collected message limit of the Reddit API.

Dataset

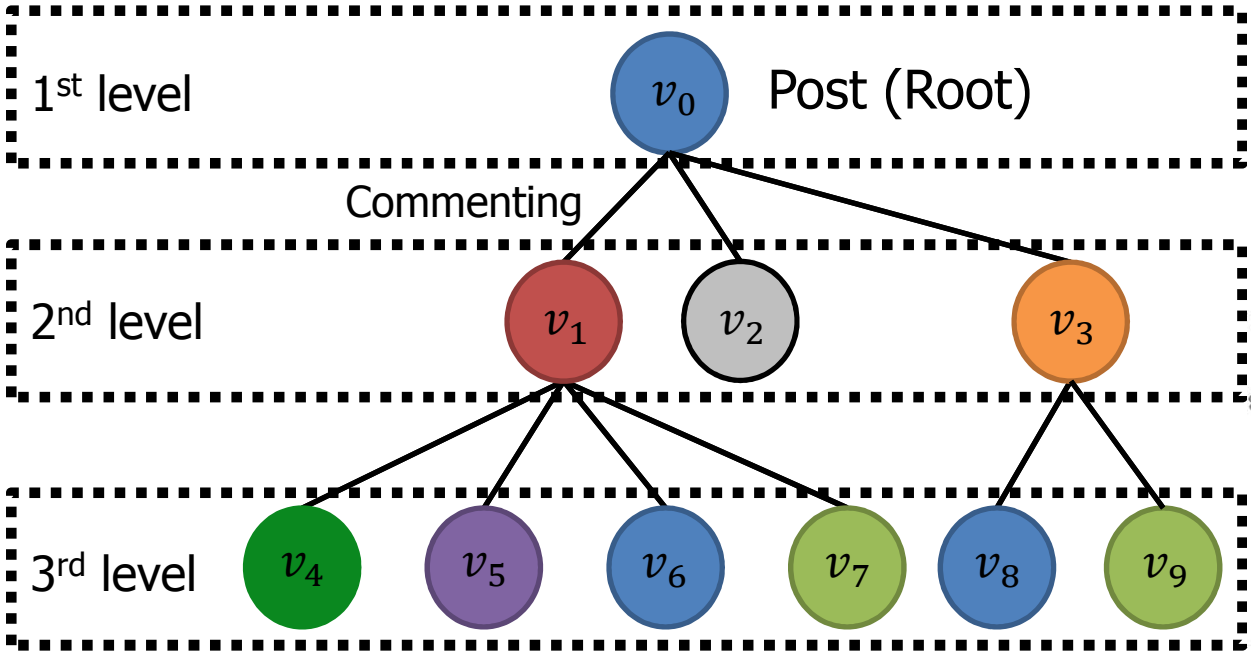
The collected dataset is stored in the DB module. We decide to choose data only from the top 100 subreddits in terms of the number of subscribers, which account for more than 60% of all subscribers (out of 378,293 subreddits, as of Oct. 22, 2014) in Reddit. We collected the dataset for 35 days from March 13 to April 18, 2014, which contains 1,016,342 posts and 18,626,530 comments, shared by 1,531,247 users. We then extracted 695,857 (68.5%) posts that each have at least one comment, and their 18,093,422 comments; posts and comments are written by 1,455,293 users. Each post contains the author id, title, subreddit id, and timestamp, while each comment contains the original post id, user id, comment text, and a parent from which the comment is generated. The parent can be a comment or a post.

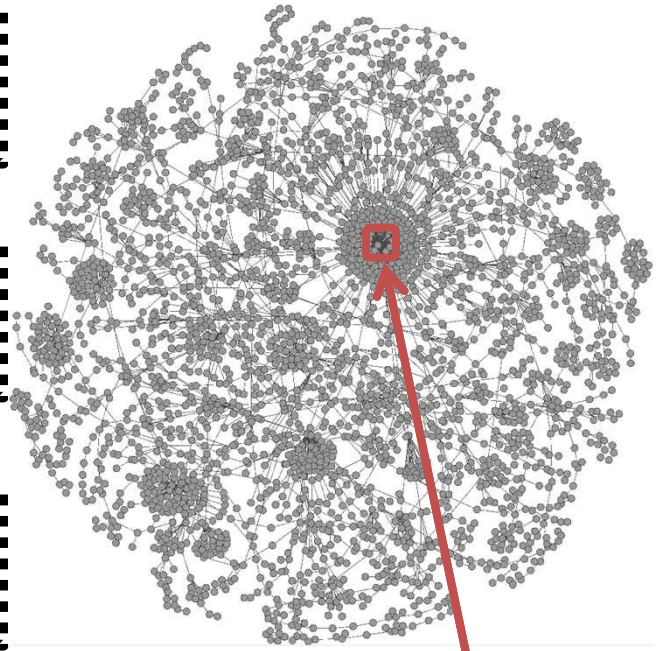
- Data Field & Description (3.3KB)
- Subreddit Information (341.7KB)
- Post Information (250.7MB)
- Comment Information (6.3GB)

You can download dataset & detailed description in here!

* Data is only available on a condition that the paper listed above is cited by your work.

Comment Tree Model



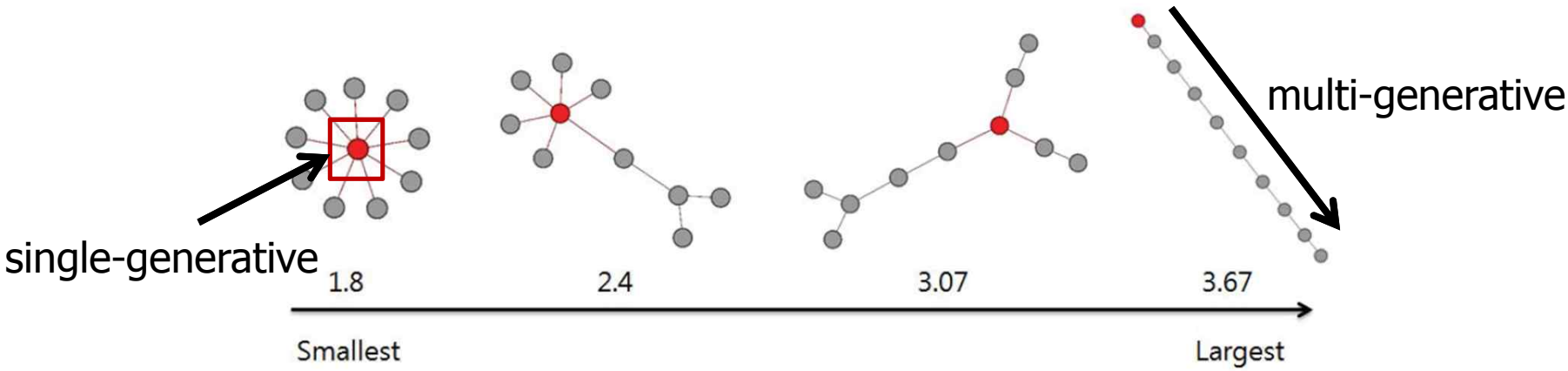


Root is here

- T = (V, E)
  - V: set of messages (a post and its all associated comments)
  - E: commenting action
- Three characteristics for comment trees
  - **Volume** – **How big** is the conversation?
  - **Responsiveness** – **How fast** do participants react in the conversation?
  - **Virality** – **How many messages elicit other messages**?

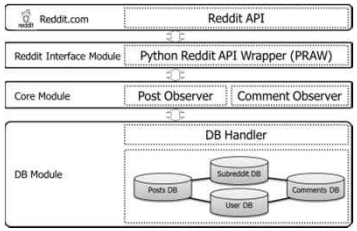
Mathematical Definition of Three Characteristics

- Volume
  - Number of nodes
- Responsiveness
  - Average of **inversed time differences** between a comment and its parent in a tree
  - Considering only the differences in range of $[\mu - 2\sigma, \mu + 2\sigma]$
- (Structural) Virality
  - Wiener Index (WI): average path length of all pairs in a tree
  - **Multi-generativity!**

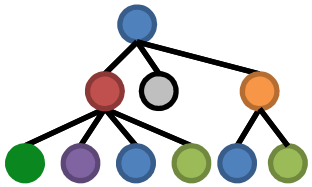


Same volume, but different virality!

Measurement Methodology



Data Collection



Comment Tree Model

Comment Tree Analysis

Content Perspective

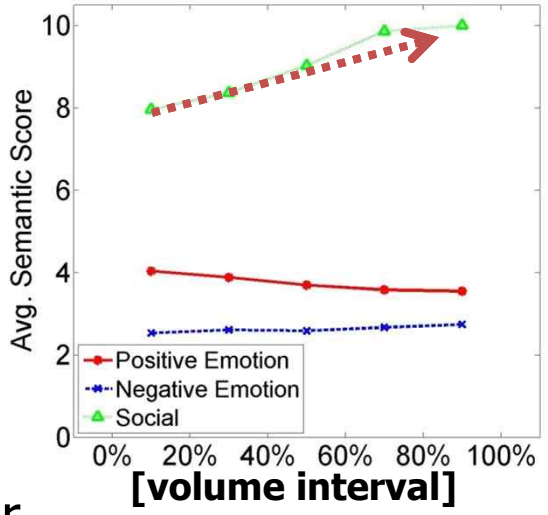
User Behavior Perspective

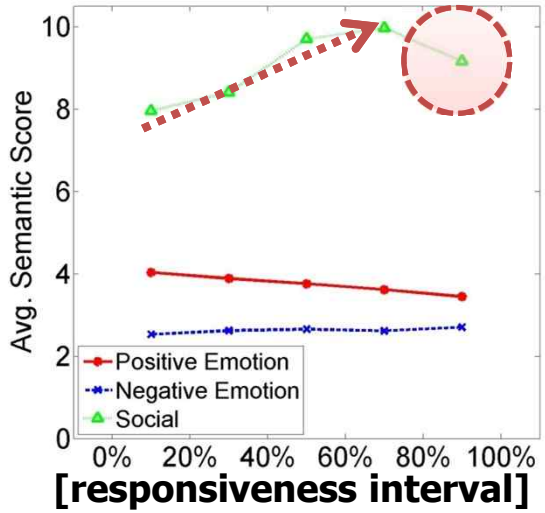
Community Analysis

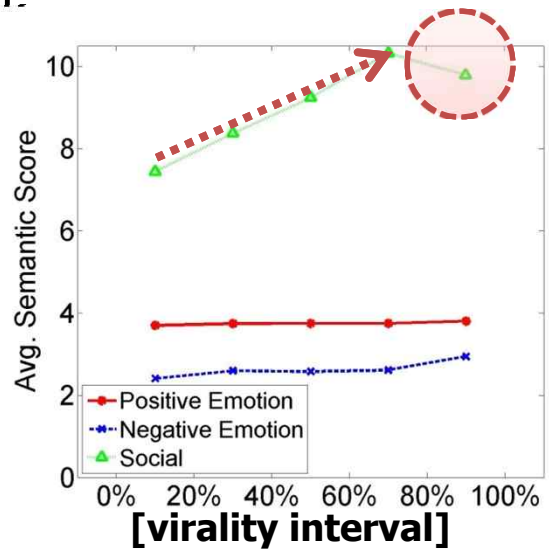
Content Perspective: Semantic Characteristics

What contents tend to be large, responsive, or viral?

- Measured by *Linguistic Inquiry Word Count* (LIWC)
  - Computing portion of categorized words
  - Social: "Mate", "Friend", "Buddy", …
  - Positive emotion: "Love", "Nice", "Sweet", …
  - Negative emotion: "Hurt", "Ugly", "Nasty", …
- Social words >> Pos. emotion > Neg. emotion
- **Social** scores become higher when trees are larger, more responsive, and more viral!
  (to a certain degree)
  - Lower in extremely responsive and viral trees
- No strong relation with **Pos./Neg.** emotion



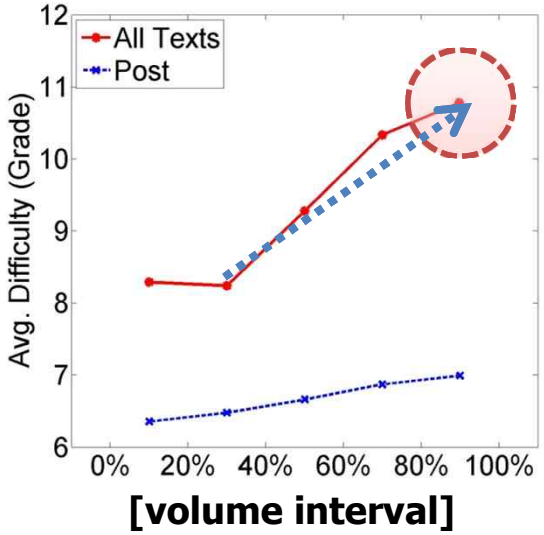




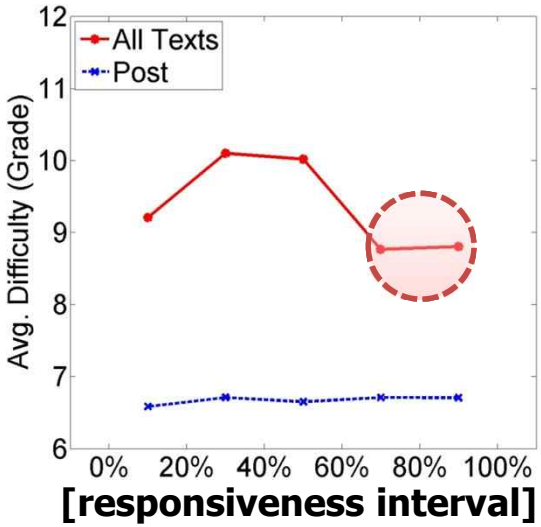
Content Perspective: Document Difficulty

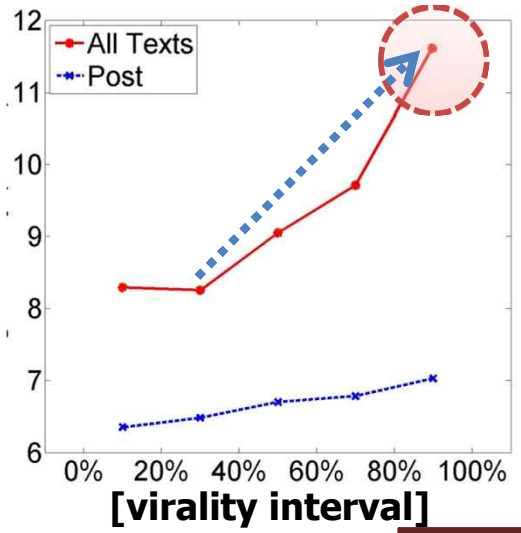
How difficulty is related to the characteristics of comment trees?

- Document Difficulty (Gunning-Fog Index)
  - Estimating **grade** of suitable student to read the text based on number of words, sentences, complex words[1)]
- Avg. difficulty of all texts >> title
  - Fewer words can be used in title



- **Volume/virality** increase**, difficulty** increases
  - More significant in viral trees
  - **Large and viral trees** tend to be **difficult**!



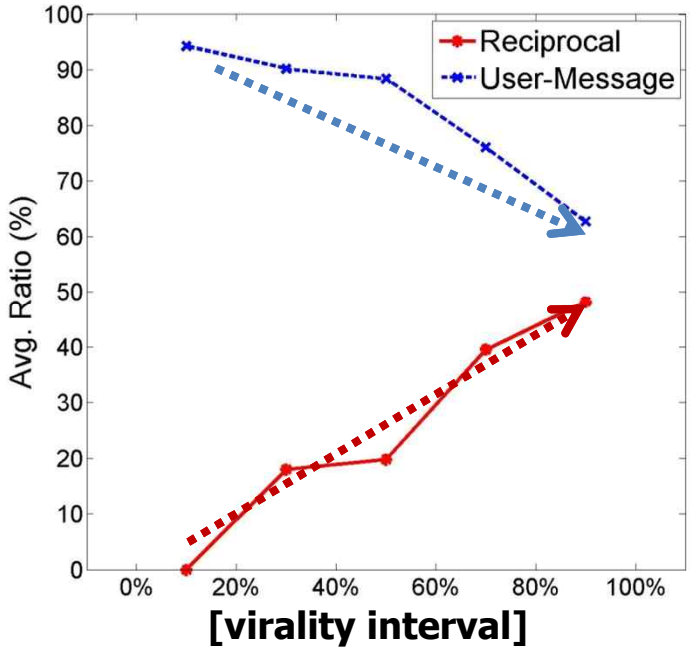


- Relatively plain texts in responsive trees

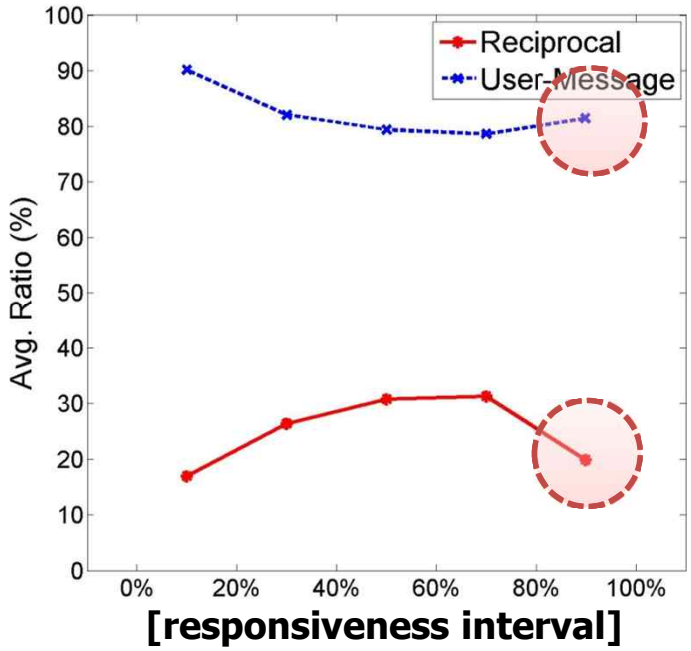
1) Complex words: words with more than 3 syllables, excluding proper nouns, familiar jargons, and compound words

User Perspective: Reciprocal & User-Message Ratio

How do users communicate in viral & responsive conversations?

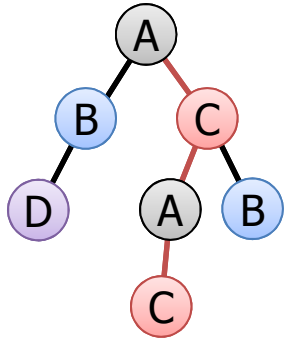


[virality interval]



[responsiveness interval]

User-Message: 4/7
Reciprocal Edge: 3/6



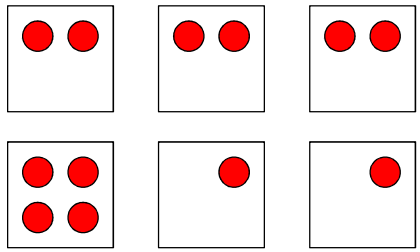
- User-Message Ratio: How many comments do users submit?
- Reciprocal Edge Ratio: How large is the portion of reciprocal communication in a tree?

- **High** reciprocal edge & **Low** user-message ratio in viral tree
  - Viral trees are formed by **reciprocal** communication from **a few of users**!
- Lower in extremely responsive trees
  - This is because that most comments are replies of the post in this case

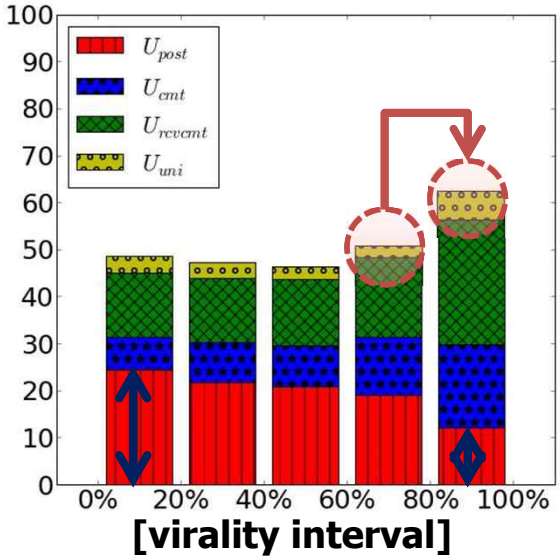
User Perspective: User Role Analysis

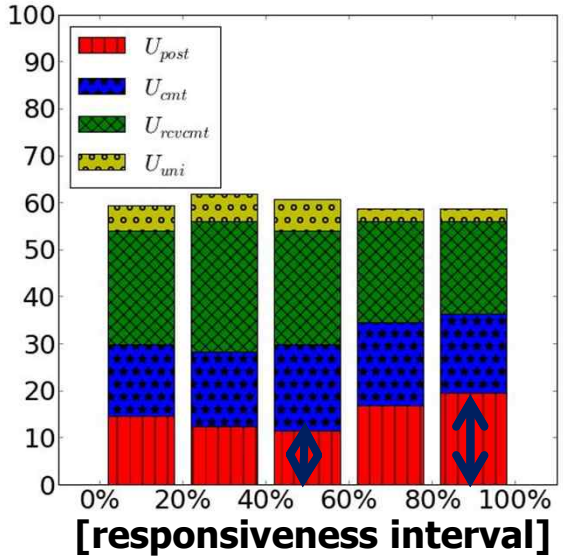
What type of users play roles in attracting other messages?

- Top 1% of users based on communication activity
  - U_{post} : by number of posts
  - U_{cmt} : by number of comments
  - U_{rcvcmt} : by number of received comments
  - U_{uni} : by **entropy** of messages across subreddits



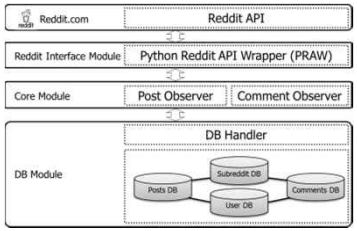
- Portion of received comments for each role type of users



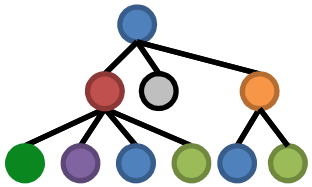


- Portion by U_{post} decreases in more viral trees, increases in more responsive trees
  - Heavy-posting users play more roles in **responsive** conversations than **viral** ones
- Portion by U_{uni} increases in the extremely viral trees
  - Number of comments by users who have broad interests could be a indicator of viral trees

Measurement Methodology



Data Collection



Comment Tree Model

Comment Tree Analysis

Content Perspective

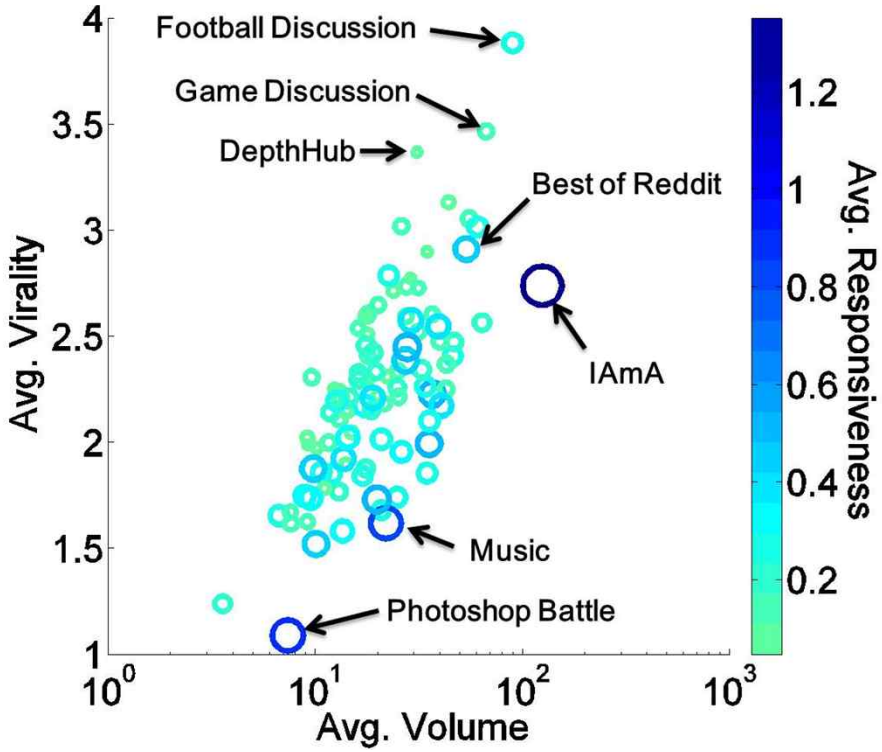
User Behavior Perspective

Community Analysis

Tree Characteristics with Topical Communities

What topical communities show large, responsive, or viral conversation?

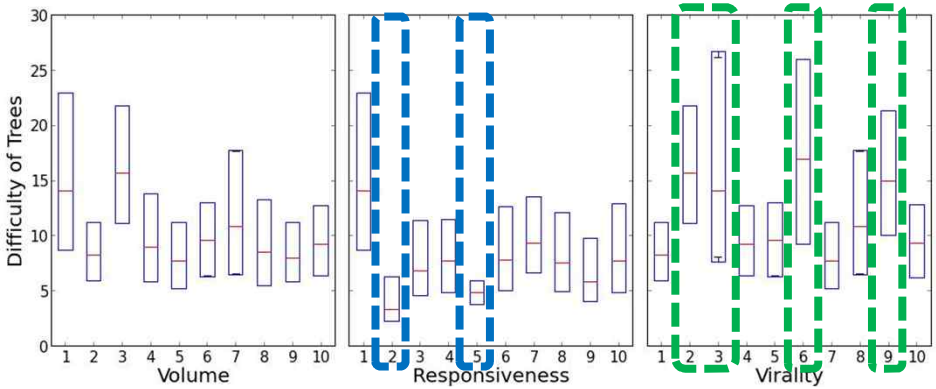
[Subreddit Map]

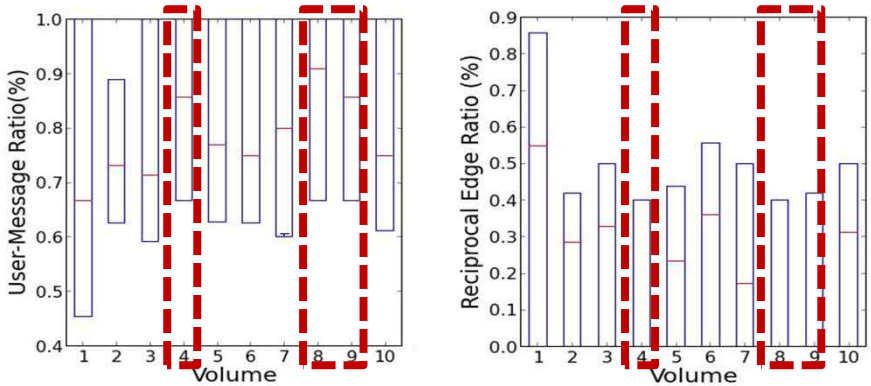


Rank	Volume	Responsiveness	Virality
1	IAmA	IAmA	Football Discussion
2	Football Discussion	Photoshop Battle	Game Discussion
3	Game Discussion	Music	DepthHub
4	Technology	Reddits Gold Mine	Android
5	Soccer	Mystery of the soda	You Should Know
6	You Should Know	Ask Reddit	The Dismal Science
7	Best of reddit	Science	Soccer
8	World News	Game of Thrones	Best of Reddit
9	TIL	FoodPorn	Frugal Living
10	Android	EarthPorn	Game Deals

- **Topical communities** show **different characteristics** of conversation
  - Large :Technology, World News, TIL — **News (or Information) sharing**
  - Responsive: Photoshop Battle, Game of Thrones, -Porn — **Multimedia-related**
  - Viral : DepthHub, The Dismal Science, Game Deals — **Discussion-related**

Content & User Behaviors in Topical Communities





Rank	Volume	Responsiveness	Virality
1	IAmA	IAmA	Football Discussion
2	Football Discussion	Photoshop Battle	Game Discussion
3	Game Discussion	Music	DepthHub
4	Technology	Reddit's Gold Mine	Android
5	Soccer	Mystery of the soda	You Should Know
6	You Should Know	Ask Reddit	The Dismal Science
7	Best of reddit	Science	Soccer
8	World News	Game of Thrones	Best of Reddit
9	TIL	FoodPorn	Frugal Living
10	Android	EarthPorn	Game Deals

- More difficult conversations in **discussion-related** subreddits
  - Difficult words are used more in discussion community!
- Relatively easier in **image-based** subreddits
- "High" user-message ratio & "Low" reciprocal edge ratio in **news-related** subreddits (Opposite patterns!)
  - REMIND: Large / viral trees tend to be generated from **reciprocal communication** from **a few of users**
  - Users in news-related subreddits **less-reciprocally** communicate!

Conclusion

- We conducted a measurement study on online conversation patterns in Reddit
  - Characterizing conversations in terms of volume, responsiveness, and virality
  - Relation of the characteristics with content & user participation behavior features

- We revealed that
  - Difficulty of content texts is an important indicator differentiating large/viral and responsive trees
  - Viral trees consist of high portion of reciprocal communication from a few of users
  - Although general tendency is kept, there are some topical communities showing distinctive characteristics

- We expect our measurement study could be the seed of researches for online conversation patterns
  - Information diffusion (word-of-mouth), Linguistics, …

Thank You!

Data & Description

http://mmlab.snu.ac.kr/traces/reddit

Contact

Daejin Choi (djchoi@mmlab.snu.ac.kr)