Hashtags on Twitter: Information Diffusion

Pawan Goyal

CSE, IITKGP

July 31, 2015

Third Reference

Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. *Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter.* In Proceedings of the 20th international conference on World wide web (WWW '11). ACM, New York, NY, USA, 695-704.

What is Information Diffusion?

Online Information Diffusion

Understanding the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have.

What is Information Diffusion?

Online Information Diffusion

Understanding the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have.

Objectives of this research

- Widespread belief that different kinds of information spread differently online.

What is Information Diffusion?

Online Information Diffusion

Understanding the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have.

Objectives of this research

- Widespread belief that different kinds of information spread differently online.
- To study this issue on Twitter, analyzing the ways in which Hashtags spread on a network defined by interactions among Twitter users.

Twitter Data and Graph Construction

- Twitter data crawled from August 2009 until January 2010.

Twitter Data and Graph Construction

- Twitter data crawled from August 2009 until January 2010.
- Collected over 3 billion messages from more than 60 million users.

Twitter Data and Graph Construction

- Twitter data crawled from August 2009 until January 2010.
- Collected over 3 billion messages from more than 60 million users.
- Graph construction via @-messages: $X \rightarrow Y$ if X directed at least 3 @-messages to Y.

Twitter Data and Graph Construction

- Twitter data crawled from August 2009 until January 2010.
- Collected over 3 billion messages from more than 60 million users.
- Graph construction via @-messages: $X \rightarrow Y$ if X directed at least 3 @-messages to Y.
- Graph size: 8.5 million non-isolated nodes, 50 million links

Twitter Data and Graph Construction

- Twitter data crawled from August 2009 until January 2010.
- Collected over 3 billion messages from more than 60 million users.
- Graph construction via @-messages: $X \rightarrow Y$ if X directed at least 3 @-messages to Y.
- Graph size: 8.5 million non-isolated nodes, 50 million links
- Studies 500 most used hashtags

Hashtag Categories

- Manually identified 8 broad categories with atleast 20 HTs in each
- Authors and 3 volunteers independently annotated each hashtag.
- Levels of agreement was high

Category	Examples
Celebrity	mj, brazilwantsjb, regis, iwantpeterfacinelli
Music	thisiswar, mj, musicmonday, pandora
Games	mafiawars, spymaster, mw2, zyngapirates
Political	tcot, glennbeck, obama, hcr
Idiom	cantlivewithout, dontyouhate, musicmonday
Sports	golf, yankees, nhl, cricket
Movies/TV	lost, glennbeck, bones, newmoon
Technology	digg, iphone, jquery, photoshop

Exposure Curve: Defining $p(k)$

Neighbor Set of X

For a given user X, the set of other users to whom X has an edge.

Exposure Curve: Defining $p(k)$

Neighbor Set of X

For a given user X, the set of other users to whom X has an edge.

When does X start mentioning a hashtag H?

How do successive exposures to H affect the probability that X will begin mentioning it?

Exposure Curve: Defining $p(k)$

Neighbor Set of X

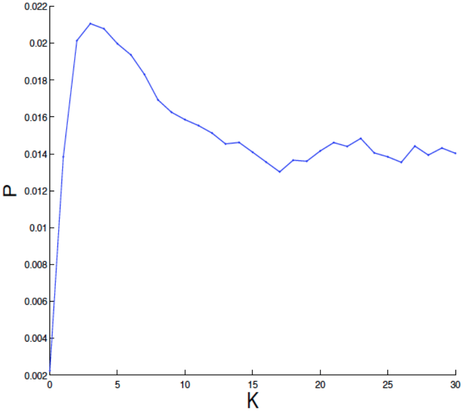
For a given user X, the set of other users to whom X has an edge.

When does X start mentioning a hashtag H?

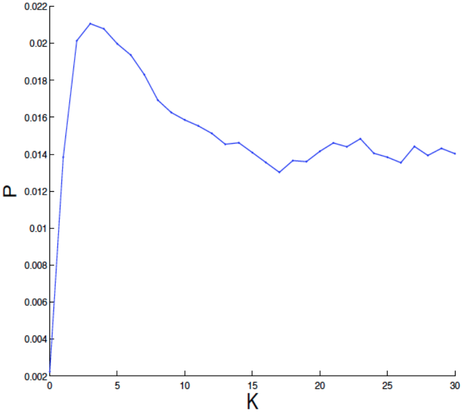
How do successive exposures to H affect the probability that X will begin mentioning it?

- Look at all users X who have not mentioned H, but for whom k neighbors have
- $p(k)$: fraction of users who adopt the hashtag *direct* after their k^{th} exposure, given that they hadn't yet adopted it.

Average Exposure Curve for 500 most-mentioned hashtags

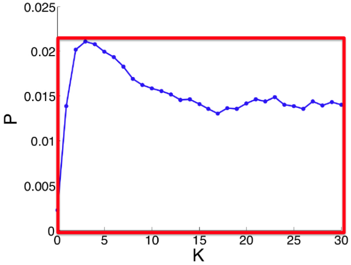


Average Exposure Curve for 500 most-mentioned hashtags



- A ramp-up to the peak value, reached relatively early ($k = 2, 3, 4$)
- Decline for larger values of k

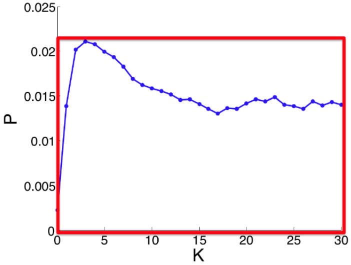
Persistence and Stickiness



Stickiness

The maximum value of $p(k)$ (probability of usage at the most effective exposure)

Persistence and Stickiness



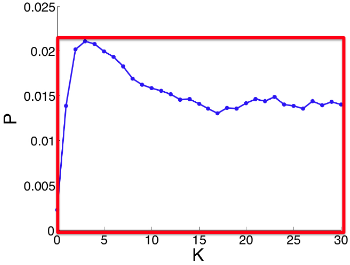
Stickiness

The maximum value of $p(k)$ (probability of usage at the most effective exposure)

Persistence

A measure of the decay of exposure curves.

Persistence and Stickiness



Stickiness

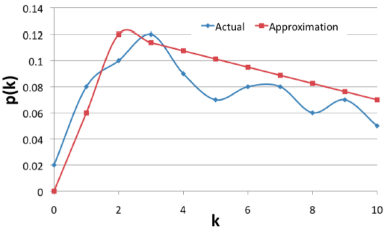
The maximum value of $p(k)$ (probability of usage at the most effective exposure)

Persistence

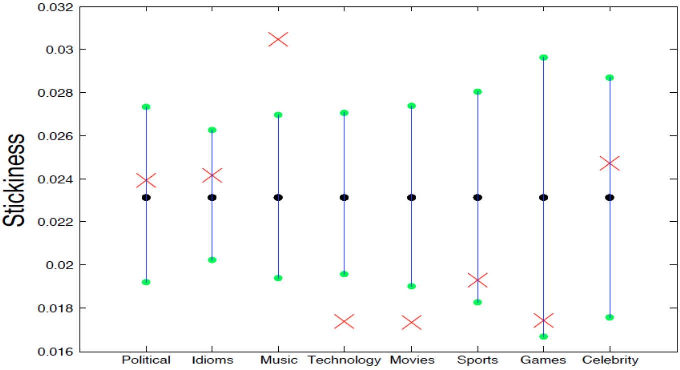
A measure of the decay of exposure curves.
The ratio of the area under the curve P and the area of the rectangle of length $max(P)$ and width $max(D(P))$.

- Are Persistence and Stickiness the adequate pair of parameters for discussing the curves' overall approximate shapes? Yes.
- Given the stickiness $M(P)$ and the persistence $F(P)$ of exposure curve P, we find an approximation $\widetilde{P}$ to P in the following way:

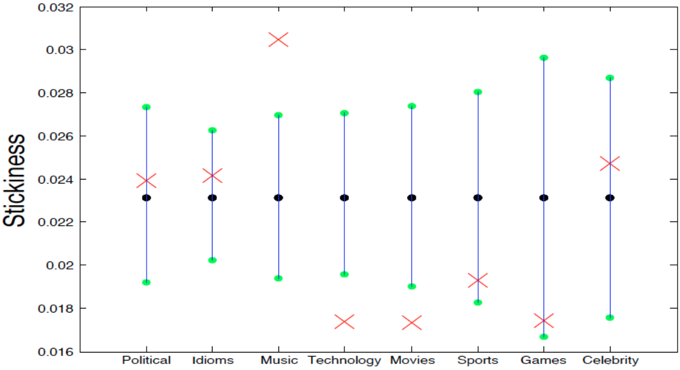
1. Let $\widetilde{P}(0) = 0$
2. Let $\widetilde{P}(2) = M(P)$
3. Now we will let $\widetilde{P}(K)$ be such that $F(\widetilde{P}) = F(P)$. This value turns out to be $\widetilde{P}(K) = \frac{M(P)*K*(2*F(P)-1)}{K-2}$
4. Make $\widetilde{P}$ piecewise linear with one line connecting the points $(0,0)$ and $(2, M(P))$, and another line connecting the points $(2, M(P))$ and $(K, \widetilde{P}(K))$.



Comparison of Hashtags based on Stickiness

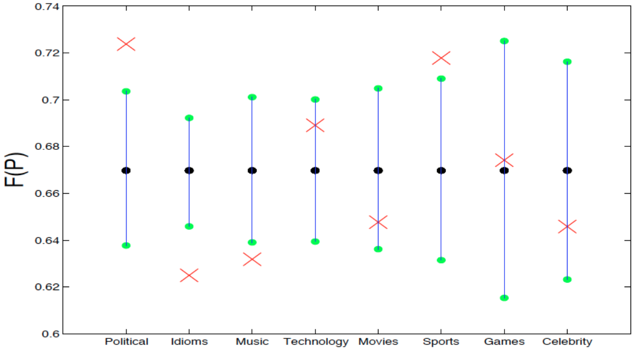


Comparison of Hashtags based on Stickiness

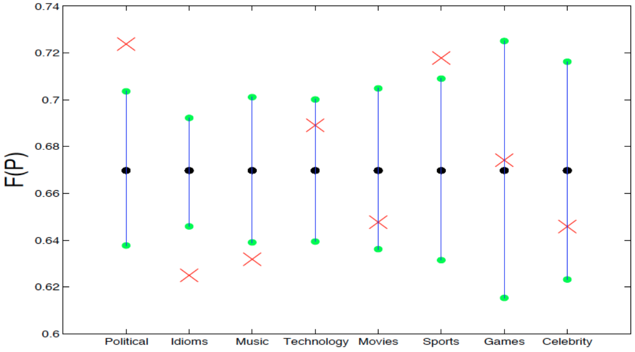


- Technology and Movies have lower stickiness than a random subset
- Music has higher stickiness than a random subset

Comparison of Hashtags based on Persistence

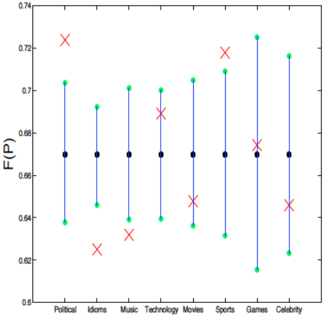


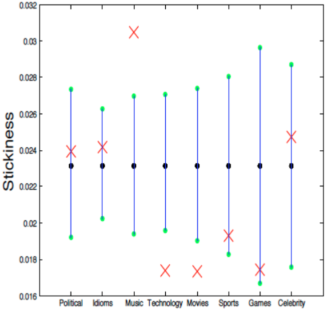
Comparison of Hashtags based on Persistence



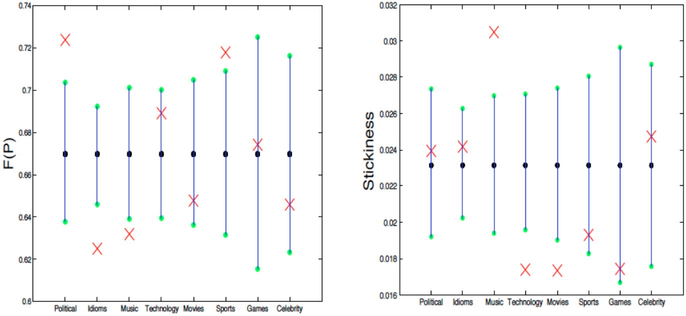
- Idioms and Music have lower persistence than a random subset of hashtags of the same size
- Politics and Sports have higher persistence than a random subset

Persistence vs. Stickiness



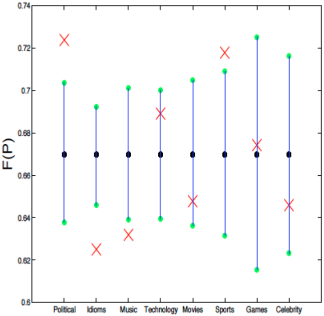


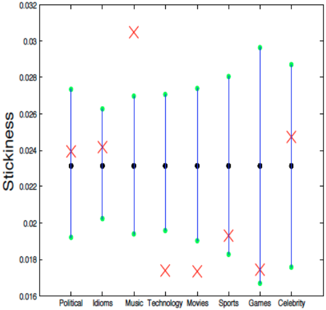
Persistence vs. Stickiness



- Idioms and Politics: Same stickiness but opposite persistence

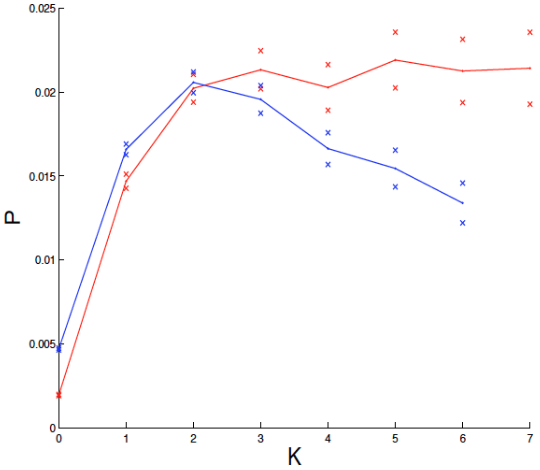
Persistence vs. Stickiness





- Idioms and Politics: Same stickiness but opposite persistence
- Music has high stickiness but low persistence
- Stickiness does not explain the diffusion well by itself

Sample curves for #cantlivewithout (blue) and #hcr (red)



Comparison of Hashtag by Mention and User Counts

Type	Mentions	Users	Mentions/User
All HTS	93,056	15,418	6.59
Political	**132,180**	**13,739**	10.17
Sports	98,234	11,329	9.97
Idioms	**99,317**	**26,319**	3.54
Movies	90,425	15,957	6.57
Celebrity	87,653	5,351	17.68
Technology	90,462	24,648	5.08
Games	123,508	15,325	6.61
Music	87,985	7,976	10.39

Table: Median Values

Comparison of Hashtag by Mention and User Counts

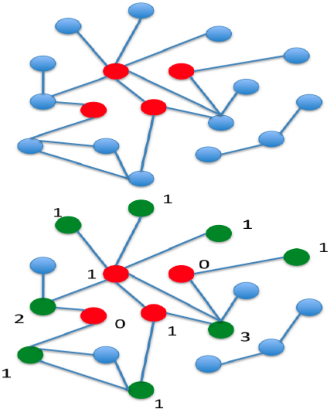
Type	Mentions	Users	Mentions/User
All HTS	93,056	15,418	6.59
Political	**132,180**	**13,739**	10.17
Sports	98,234	11,329	9.97
Idioms	**99,317**	**26,319**	3.54
Movies	90,425	15,957	6.57
Celebrity	87,653	5,351	17.68
Technology	90,462	24,648	5.08
Games	123,508	15,325	6.61
Music	87,985	7,976	10.39

Table: Median Values

Political and Idioms are among the most mentioned, but Idioms are used by twice the number of people that use Politics

The Structure of Initial Sets

- Let G_m be the subgraph induced by the first m users of a given hashtag.
- Let the *border* of G_m be the set of nodes not in G_m with at least one edge to a node in G_m.
- Let the *internal degree* of a node in G_m be the number of neighbors it has in G_m.
- Let the *entering degree* of a node in the border of G_m be the number of neighbors it has in G_m



Structure Comparison for Political Hashtags (G_{500})

Type	Internal Degree	Triangle Num	Entering Deg.	Border Nodes
All HTS	1.41	384	1.24	13425
Political	2.55	935	1.41	12879
Upper Error Bar	1.82	653	1.32	15838
Lower Error Bar	1.00	112	1.16	11016

Structure Comparison for Political Hashtags (G_{500})

Type	Internal Degree	Triangle Num	Entering Deg.	Border Nodes
All HTS	1.41	384	1.24	13425
Political	2.55	935	1.41	12879
Upper Error Bar	1.82	653	1.32	15838
Lower Error Bar	1.00	112	1.16	11016

- The early adopters of a political hashtag message more with each other, create more triangles, and have a border of people with more links into the early adopter set.