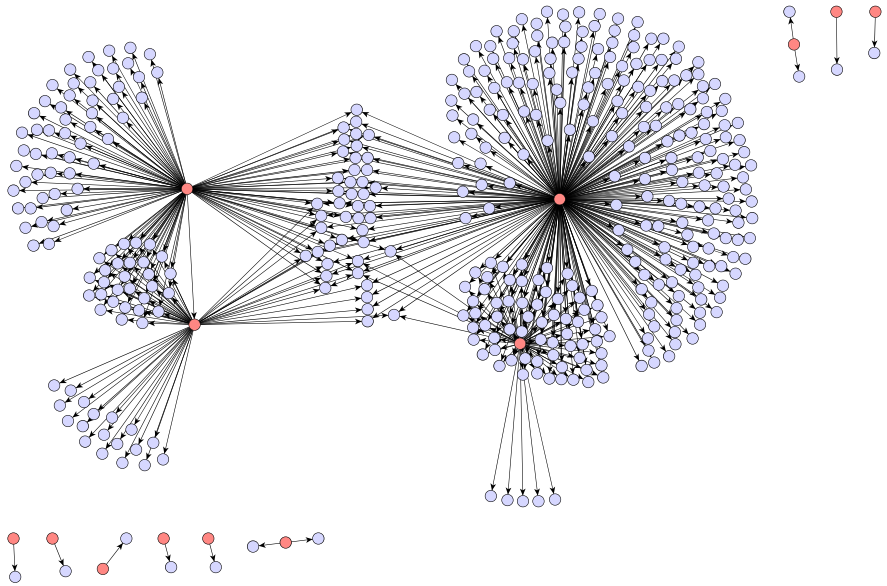
Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter

Daniel M. Romero, Brendan Meeder, and Jon Kleinberg

Cornell University and Carnegie Mellon University

Online 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.



[Leskovec-Adamic-Huberman 2006]

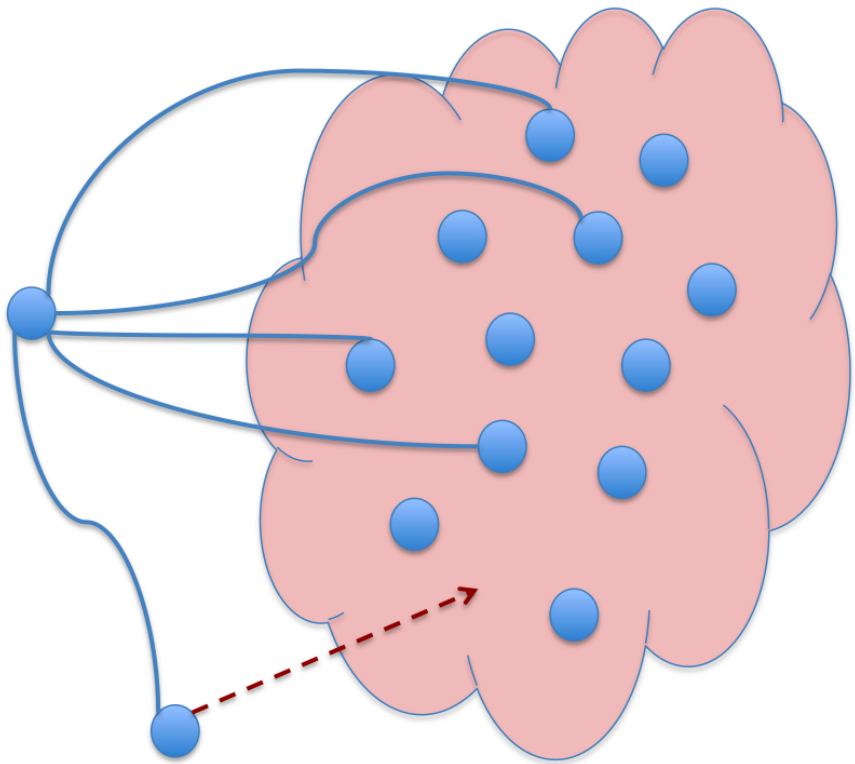
- Previous Work: Finding properties that generalize across different topics and different types of information.
- Current Work: Finding sources of variation across different topics and types of information.

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 sent at least 3 @-messages to Y.
- Graph Size: 8.5 million non-isolated nodes and 50 million links.
- Studied 500 most used hashtags

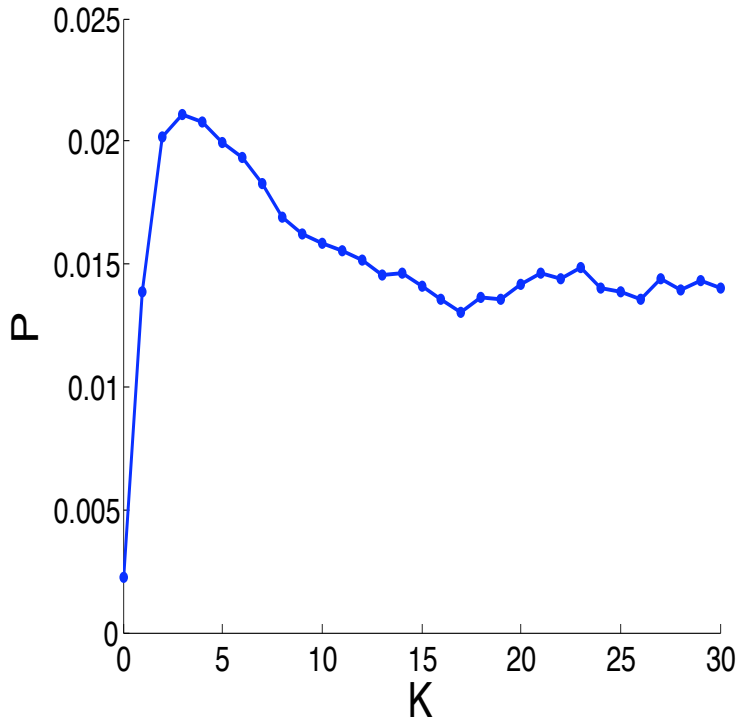
Exposure Curves Definition

- For user X, we call the set of other users to whom X has an edge the *neighbor set* of X.
- Fix a hashtag H.
- Define $E(k)$ as all users X who have not yet mentioned H, but for whom k neighbors have.
- Define $p(k)$ to be the fraction users in $E(k)$ who mention H before a $(k+1)^{st}$ neighbor does.
- We call $p(k)$ the exposure curve of H.



[Cosley, et al. 2010]

Average Exposure Curves

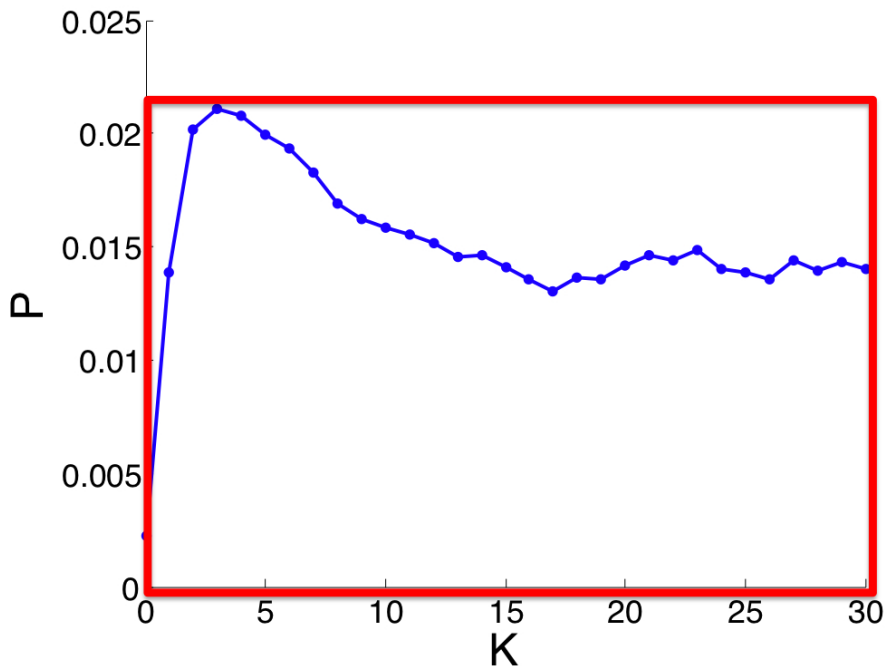


Average exposure curve for the top 500 hashtags.

What are the most important aspects of the shape of exposure curves?
Curve reaches peak fast, decreases after.

Persistence and Stickiness

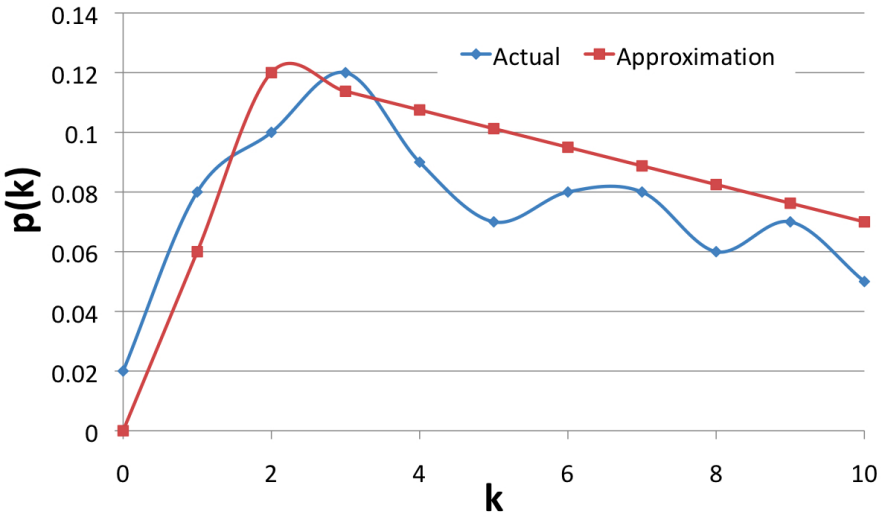
- *Persistence* of P is the ratio of the area under the curve P and the area of the rectangle of length $max(P)$ and width $max(D(P))$, where $D(P)$ is the domain of P.
- Persistence measures the decay of exposure curves.
- *Stickiness* of P is $max(P)$.
- Stickiness is the probability of usage at the most effective exposure.



Approximating Exposure Curves via Stickiness and Persistence

- 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 $(K, \widetilde{P}(K))$.

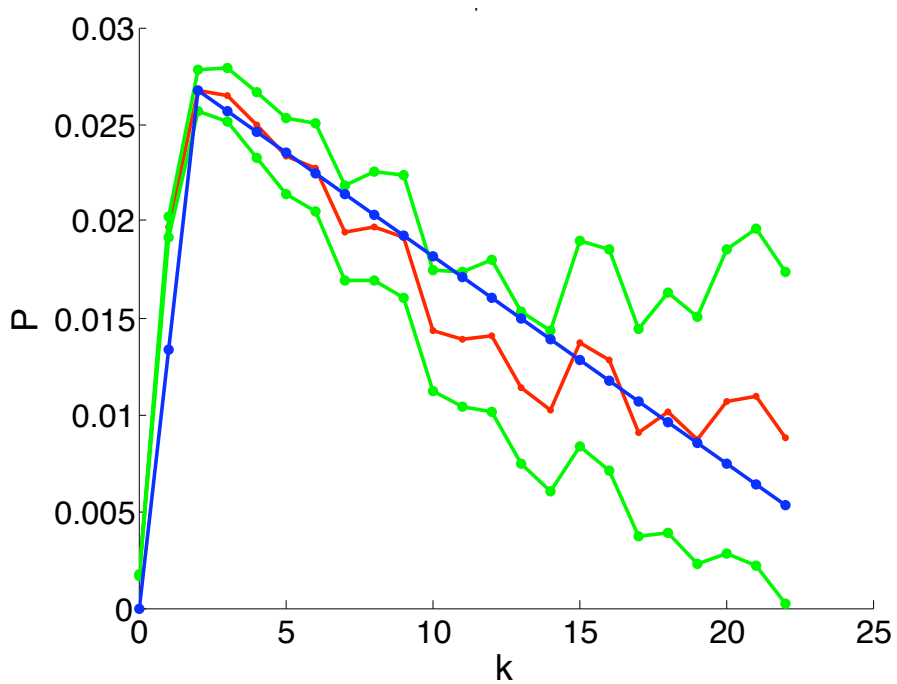


Approximating Exposure Curves via Stickiness and Persistence

- Approximation error defined as the mean absolute error:

$$E(P,\widetilde{P}) = \frac{1}{K}\sum_{k=0}^{K}\left|(P(k)-\widetilde{P}(k))\right|$$

- Average approximation error is 0.0050
- Average error of based on the exposure curves' confidence intervals is 0.0056.
- Approximation error is about the same as error based on confidence intervals



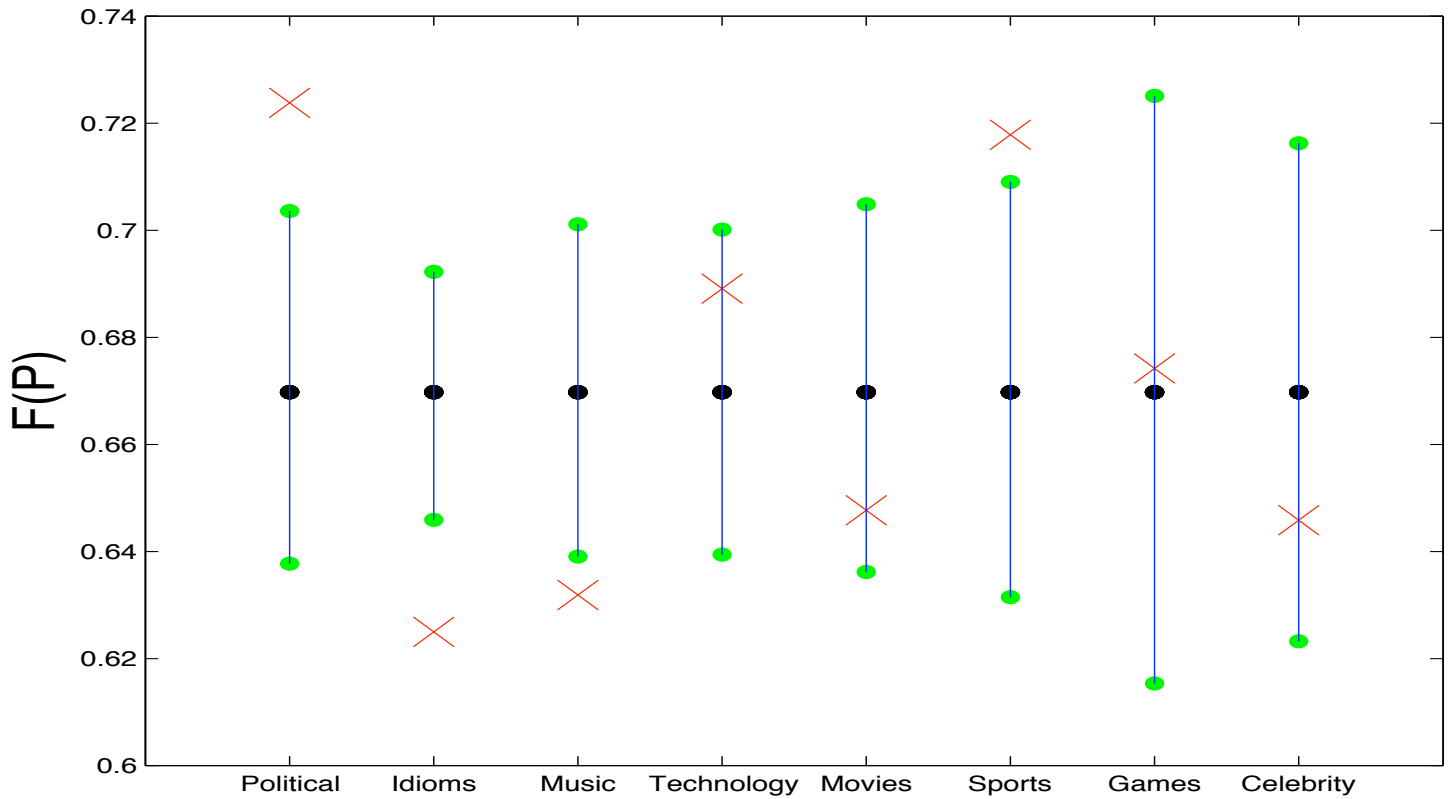
Hashtags Categories Based on Top 500

- Manually identified 8 broad categories with at least 20 HTs in each
- Then authors and 3 volunteers independently annotated each hashtag.

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

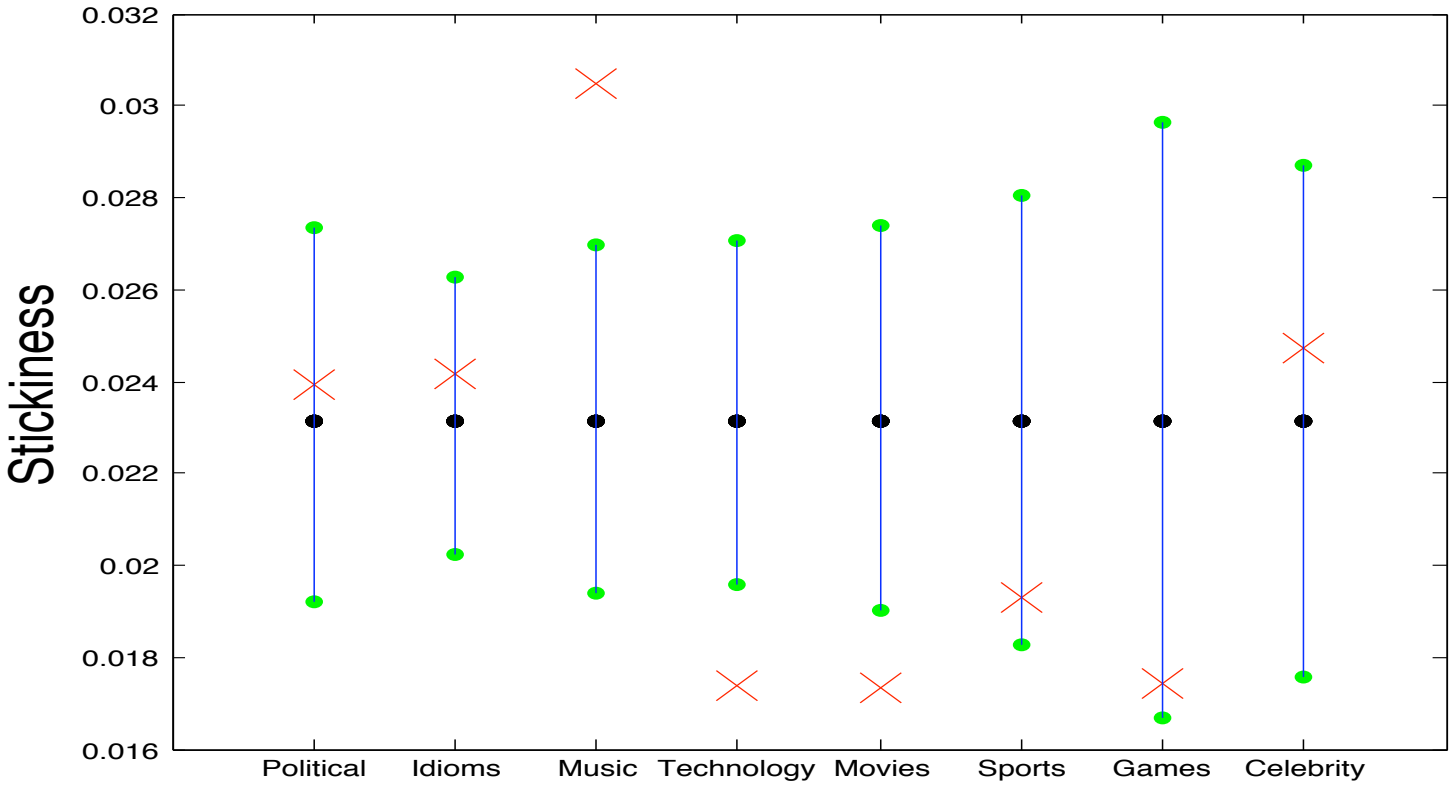
- Level of agreement was high.
- Results are essentially identical when based on the authors' annotations, volunteers' annotations, or the intersection.

Comparison of Hashtags Based on Persistence



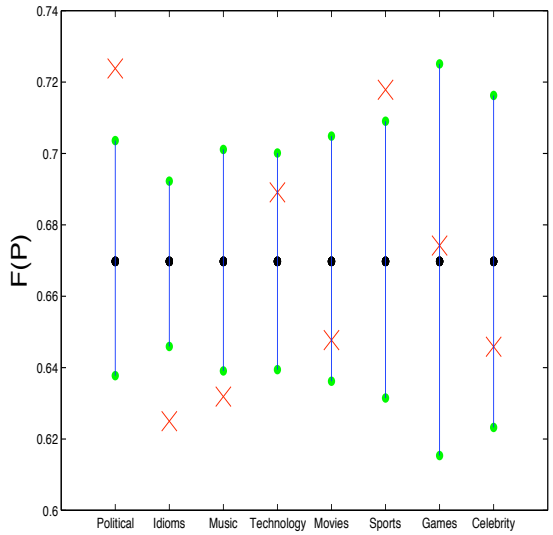
- Idioms and Music have lower persistence than that of a random subset of hashtags of the same size.
- Politics and Sports have higher persistence than that of a random subset of hashtags of the same size.

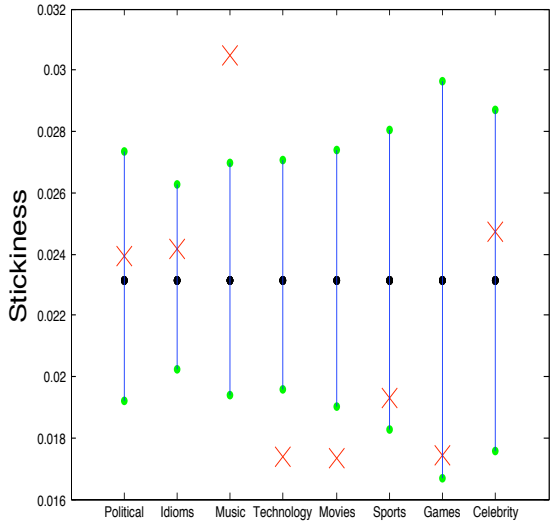
Comparison of Hashtags Based on Stickiness



- Technology and Movies have lower stickiness than that of a random subset of hashtags of the same size.
- Music has higher stickiness than that of a random subset of hashtags of the same size.

Persistence vs. Stickiness





- Idioms and Politics: Same stickiness but opposite persistence – Complex Contagion [Centola-Macy 2007].
- Music has high stickiness but low persistence
- Stickiness does not explain the diffusion mechanism well by itself.

Comparison of Hashtag by Mention and User Counts

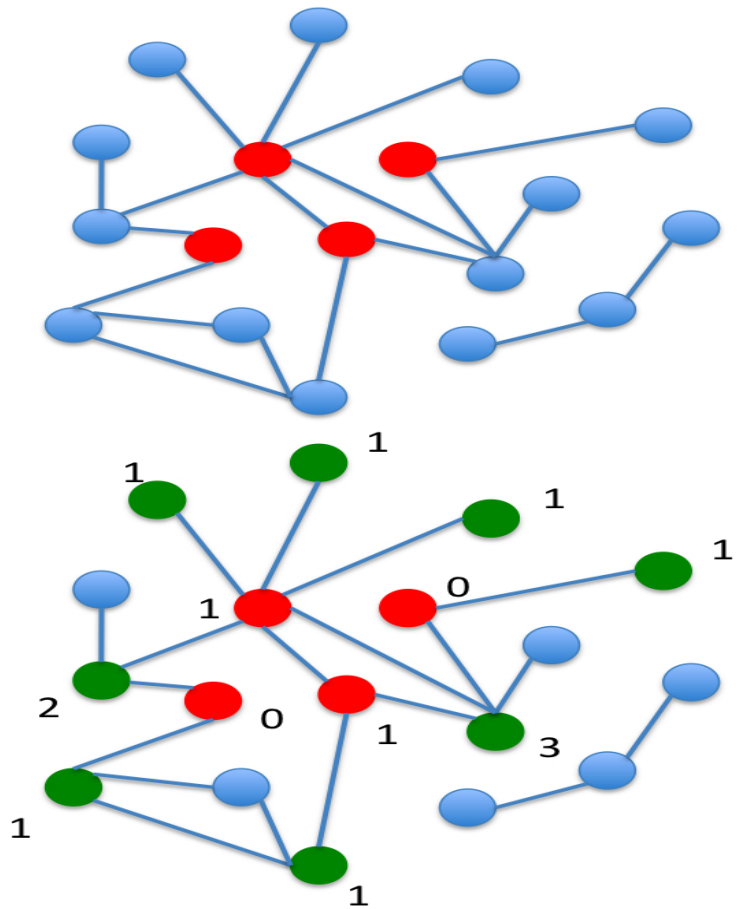
Type	Mentions	Users	Mentions/User
All HTS	93,056	15,418	6.59
Political	**132,180**	**13,739**	10.17
Sports	98,234	11,329	9.97
Idioms	**99,317**	**26,319**	3.54
Movies	90,425	15,957	6.57
Celebrity	87,653	5,351	17.68
Technology	90,462	24,648	5.08
Games	123,508	15,325	6.61
Music	87,985	7,976	10.39

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



G_{500} Structure Comparison for Political Hashtags

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

Subgraphs G_m corresponding to political hashtags exhibit the most significant structural differences from the average:

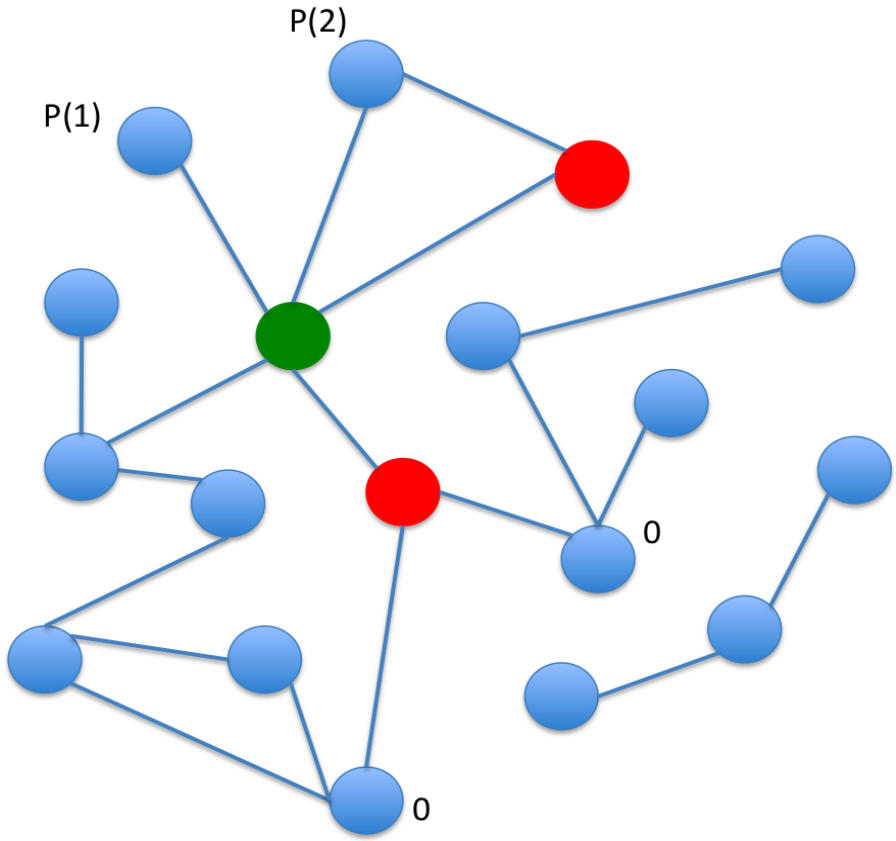
The early adopters of a political hashtag message more with each other (Consistent with McAdam's theories [McAdam 1986, 1988]), create more triangles, and have a border of people with more links into the early adopter set.

The Simulated Model

Do initial user graphs and exposure curves go together? Do they depend on each other to spread hashtags?

Fix exposure curve $p(k)$, graph $G = (V, E)$, and initially activated node I. For each iteration t (Starting with $t = 0$):

1. A_t = Currently active nodes.
2. $N_t \subseteq A_t$ = Newly active nodes. (Initially $A_0 = N_0 = I$)
3. Nodes in N_t activate (infect) each of their inactive neighbors u with probability $p(k)$ where k is the number of infected neighbors of u.



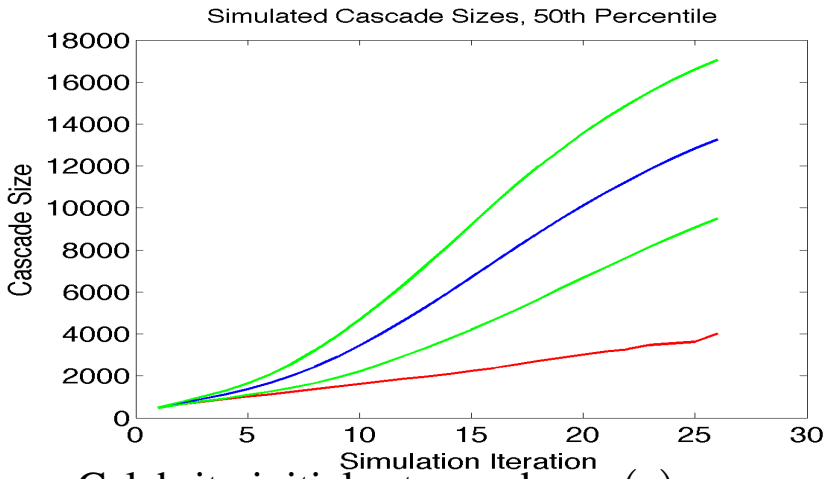
A_0 = Red

A_1 = Red $\cup$ Green

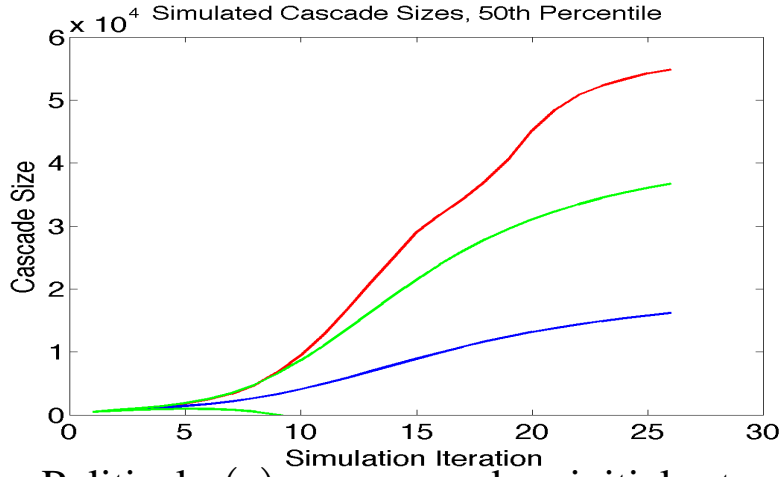
N_1 = Green

Simulation Results

- Political and Idioms: Both $p(k)$ curves and initial sets produce larger cascades when paired within the same category
- Celebrities and Games: Initial sets produce smaller cascades when paired with $p(k)$ curves from the same category
- Music: Initial sets produce larger cascades with $p(k)$ curves from the same category. And $p(k)$ curves produce smaller cascades with initial sets from the same category
- Movies, Sports, and Technology: No statistically significant trend



Celebrity initial sets, random $p(x)$ curves.



Political $p(x)$ curves, random initial sets.

Conclusion

- Hashtags of different types and topics exhibit different mechanics of spread.
- Differences in spread can be analyzed in terms of the probabilities of adoption after repeated exposures.
- We observe variations of these probabilities in magnitude (stickiness) as well as rate of decay (persistence).
- The adoption of politically controversial hashtags is especially affected by multiple repeated exposures which provides a validation the complex contagion principle.
- The graph induced by the initial set of adopters of political hashtags have a particular structure that agrees with previous sociological theories.
- We begin to understand the relationship between the exposure curves and graph structure through simulations (see the paper).