MIS Quarterly

RESEARCH ARTICLE

USING RETWEETS WHEN SHAPING OUR ONLINE PERSONA: TOPIC MODELING APPROACH

Hilah Geva and Gal Oestreicher-Singer
The Coller School of Management, Tel-Aviv University,
Tel Aviv, ISRAEL {hilahlev@mail.tau.ac.il} {galos@post.tau.ac.il}

Maytal Saar-Tsechansky
McCombs School of Business, University of Texas at Austin,
Austin, TX 79712 U.S.A. {maytal@mail.utexas.edu}

Appendix A

Robustness Analysis Using an Additional Data Set

All analyses in this paper were done on a data set that was collected in November 2016 (going back 6 months). When analyzing this data set, we assumed that tweets appear in a chronological order. However, in February 2016, Twitter changed the home timeline layout to highlight certain tweets, in which the user is likely to be interested. This means that during our collection period, users were not necessarily presented with tweets in strict chronological order.

We control for this potential source of bias by rerunning our main analysis on a previously collected data set. This data set includes tweets of 2,435 core users (and their followings) from September 2015 (going back 6 six), that is, prior to the policy change. All results obtained for this data set are indeed similar in direction, magnitude, and significance to those obtained from the 2016 data set.

While the 2015 data set circumvents bias related to Twitter's policy change, it has a different limitation: The filters used to choose the core users were somewhat stricter than those used in the 2016 collection, as follows:

(1) To be included in our 2015 data set, a user had to have posted at least 200 retweets and self-tweets in the 6-month period. Additionally, we filtered out users with fewer than 15 retweets or fewer than 15 self-tweets in each of the three months prior to the date of collection.

(2) We filtered out users with exceptionally high and low (top or bottom 15%) numbers of followers or followings.

While each dataset suffers from its own limitations, these limitations do not overlap, enabling us to suggest that the consistency of our results between the two complementary data sets offers robustness to our results.

Tables A1 to A6 present the main results for H1, H2, and H3 for the 2015 data set.

Table A1. H1: Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method for Core Users Data Set

	***Th* = 1**	***Th* = 2**	***Th* = 5**	***Th* = 10**
Mean of number of topics added via retweeted persona	1.99	1.88	1.71	1.54
Mean of number of topics added via random retweeted persona	4.90	4.60	3.71	2.93
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table A2. H1: Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method Core Users Data Set

	***Th* = 0.8**	***Th* = 0.85**	***Th* = 0.9**	***Th* = 0.95**
Mean of number of topics added via retweeted persona	1.47	1.55	1.64	1.77
Mean of number of topics added via random retweeted persona	2.77	3.05	3.39	3.94
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table A3. H2: Mean J-S Divergence Between Self Tweets and Full (Random) Personas for Core Users Data Set

	Mean J-S Divergence
Self-tweet and full persona	0.057
Self-tweets and random full persona	0.073
Wilcoxon signed-rank test	$p < 0.001$

Table A4. H3: Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method for Combined Data Set

	Experts			
	***Th* = 1**	***Th* = 2**	***Th* = 5**	***Th* = 10**
Mean of number of topics added via retweeted persona	1.49	1.13	0.68	0.49
Mean of number of topics added via random retweeted persona	4.61	3.78	2.41	1.52
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
	Core Users			
	Th = 1	*Th* = 2	*Th* = 5	*Th=10*
Mean of number of topics added via retweeted persona	2.15	1.90	1.59	1.38
Mean of number of topics added via random retweeted persona	5.7963	5.08994	3.60329	2.49
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table A5. H3: Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method Combined Data Set

	Experts			
	***Th* = 0.8**	***Th* = 0.85**	***Th* = 0.9**	***Th* = 0.95**
Mean of number of topics added via retweeted persona	0.55	0.69	0.98	1.35
Mean of number of topics added via random retweeted persona	1.93	2.44	3.11	4.07
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001
	Core Users			
	***Th* = 0.8**	***Th* = 0.85**	***Th* = 0.9**	***Th* = 0.95**
Mean of number of topics added via retweeted persona	1.29	1.36	1.45	1.63
Mean of number of topics added via random retweeted persona	2.15	2.46	2.96	3.87
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Table A.6. H3: Mean J-S Divergence Between Self Tweets and Full (Random) Personas for Combined Data Set

	Mean J-S Divergence	
	Core Uers	***Bloggers***
Self-tweet and full persona	0.054	0.012
Self-tweets and random full persona	0.067	0.032
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001

Appendix B

Recreation of Home Timelines

On Twitter, the home timeline of a user *u* is a stream of all tweets (self-tweets and retweets) posted by all the users that *u* follows, sorted approximately in reverse chronological order, including promoted tweets that originate from Twitter. One should note that, as a general rule, retweeting does not cause tweets to reappear in a user's home timeline or to change their relative position in the timeline. For example, if *u* follows user *a*, and *a* posts a tweet, this tweet will appear only once in user *u's* home timeline regardless of whether it was retweeted by other users *u* follows.

However, when collecting data from the REST API, it is not possible to retrieve a user's full home timeline as such. Instead, we only observe each user's *user timeline*: the tweets and retweets that he or she has posted. Consequently, the recreation of the home timeline of each core user was done in two steps: First, we combined all tweets posted by the core user's followings into one timeline and sorted them by creation date. Second, since retweets do not cause tweets to reappear in a user's home timeline, we filtered out duplicate tweets. For example, if *u* follows both *a* and *b*, who both retweet tweet *t*, in reality this tweet will appear only once in *u's* home timeline, attributed to the user who retweeted first (let us assume that it is *a*). However, when we combine the user timelines of *a* and *b*, tweet *t* appears twice (once from *a* and once from *b*), so it is necessary to filter *b*'s retweet out of the timeline. These two steps produce a timeline that closely approximates the actual home timeline of user *u.*

Note that some followings had their privacy settings set to private during the data collection period or had their accounts suspended, meaning that we could not collect their data. This means that our recreated timelines missed some incoming tweets. However, we estimate that the number of followings affected should not exceed 10% of the followings in our data set.

Appendix C

Data Collection for Expert Users

The selection of expert users was done under the assumption that individuals who blog for prominent blog sites and also have Twitter accounts are particularly likely to use Twitter as a personal branding tool. We therefore manually gathered lists, from the Twitter pages of 12 blogging websites, that contained the Twitter accounts of the bloggers who contribute to those sites. Table C1 presents the URLs of the blogs, their Twitter pages, and the specific list pages from which we composed the set of expert users.

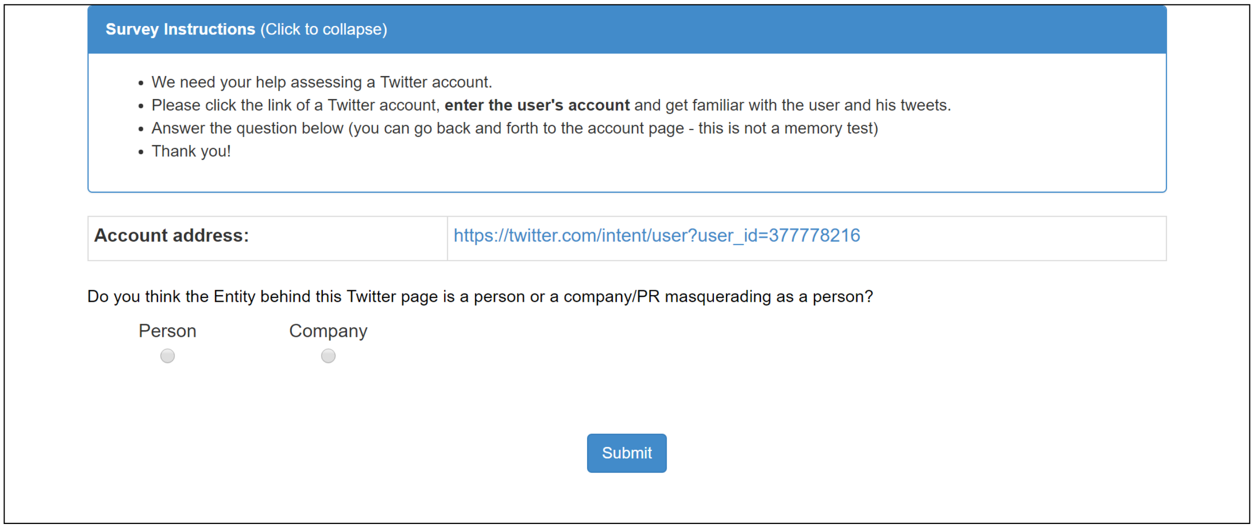
Table C1. Blog URLs

Blog Name	Website	Twitter Page	Lists from the Twitter Page
Huffington Post	http://www.huffingtonpost.com/	https://twitter.com/HuffingtonPost	https://twitter.com/HuffingtonPost/lists/tech-politics-bloggers/members https://twitter.com/HuffingtonPost/lists/huffposters-2/members https://twitter.com/HuffingtonPost/lists/bloggers/members
Business Insider	http://www.businessinsider.com/	https://twitter.com/businessinsider	https://twitter.com/businessinsider/lists/bi-editors-reporters/members
Mashable	http://mashable.com/	https://twitter.com/mashable/	https://twitter.com/mashable/lists/mashable-staff-24/members
Gizmodo	http://gizmodo.com/	https://twitter.com/Gizmodo	https://twitter.com/Gizmodo/lists/writers/members https://twitter.com/Gizmodo/lists/gizmodostaff/members
Lifehacker	http://lifehacker.com/	https://twitter.com/lifehacker	https://twitter.com/lifehacker/lists/lifehacker/members
Gawker	http://gawker.com/	https://twitter.com/Gawker	https://twitter.com/Gawker/lists/writers/members
The Daily Beast	http://www.thedailybeast.com/	https://twitter.com/thedailybeast	https://twitter.com/thedailybeast/lists/the-daily-beast-staff/members
Techcrunch	http://techcrunch.com/	https://twitter.com/TechCrunch	https://twitter.com/TechCrunch/lists/writers/members
Jezebel	http://jezebel.com/	https://twitter.com/Jezebel/	https://twitter.com/Jezebel/lists/writers/members https://twitter.com/Jezebel/lists/jezebel-guide/members
The next web	http://thenextweb.com/	https://twitter.com/TheNextWeb	https://twitter.com/TheNextWeb/lists/tnw-team/members
Epicurious	http://www.epicurious.com/	https://twitter.com/epicurious	https://twitter.com/epicurious/lists/epicurious-editors-2/members
NYT Food	http://www.nytimes.com/pages/dining/index.html	https://twitter.com/nytfood	https://twitter.com/nytfood/lists/foodies/members

Appendix D

MTurk Survey to Determine Whether Experts' Accounts Represent Actual People or Services

To make sure our experts' (bloggers') accounts corresponded to real-life individuals rather than to services or products, we ran the following survey on Amazon Mechanical Turk (see Figure D1), asking workers to tell us whether each blogger's account reflected a company or an actual person. Each account was graded by three unique Turkers. We say an account represents an actual person if at least two out of the three Turkers marked it as such.



Survey Instructions (Click to collapse)

- We need your help assessing a Twitter account.
- Please click the link of a Twitter account, **enter the user's account** and get familiar with the user and his tweets.
- Answer the question below (you can go back and forth to the account page - this is not a memory test)
- Thank you!

Account address:	https://twitter.com/intent/user?user_id=377778216

Do you think the Entity behind this Twitter page is a person or a company/PR masquerading as a person?

Person ○ Company ○

Figure 1. Population Model (Simple Model)

Appendix E

Identification

To clarify our identification strategy, we present a diagram that visually portrays our strategy. Let's assume Dan's home timeline (the tweets he sees) consists of the following 15 tweets:

Tweet 1 "it's like i always say: cleveland is bad"
Tweet 2 "musicianship: what a load of fascist malarkey"
Tweet 3 I just witnessed a DRIVER of a CAR run a Red Light. Time to fire off an Op-Ed to the @chicagotribune calling ALL drivers scofflaws! #bikeCHI
Tweet 4 Carrie just struck a bowl & said: "might be an ugly bowl, but it sounds good"
Tweet 5 "Ever sine you told me you saw that Diners, Drive-ins, and Dives guy in NYC, I just won't watch his show any more"#DadTime
Tweet 6 Thinking a lot about music that exists in spaces where it's left unconsidered. How & why it's made and by whom?
Tweet 7 When the servers try and tell you a joke saying you sound like an owl and you actually start crying bc you thought someone was being mean
Tweet 8 Offensive line: TAKE CARE OF LAMAR
Tweet 9 Plot twist: the white girl isn't drinking a PSL
Tweet 10 Update: just got reprimanded for getting on tinder. This is why I have a privacy screen. I need my replacement asap
Tweet 11 Tinder in NYC is really depressing because everyone is beautiful and you're just irrelevant
Tweet 12 Get snaps of the inside of frat life is very entertaining. Dance pledges, dance
Tweet 13 When they said I could do better, they were damn right
Tweet 14 Under U.S. law Hillary literally is disqualified from becoming president... https://t.co/N7h4mV88h5
Tweet 15 New uniforms to honor POW/MIA soldiers & all veterans who served our great nation @ our Military Appreciation Game-… https://t.co/GPs6t8xRVa

Now, let us assume that from these 15 tweets Dan retweeted tweets 4, 7, and 15. In this case his **retweeted persona** will be

Tweet 4 Carrie just struck a bowl & said: "might be an ugly bowl, but it sounds good"
Tweet 7 When the servers try and tell you a joke saying you sound like an owl and you actually start crying bc you thought someone was being mean
Tweet 15 New uniforms to honor POW/MIA soldiers & all veterans who served our great nation @ our Military Appreciation Game-… https://t.co/GPs6t8xRVa

To build Dan's random retweeted persona we randomly sample three tweets from Dan's feed (as we explain, we sample the exact number of retweets that the user actually posted). Let's say we randomly sampled tweets 3, 7, and 10. This means that Dan's **random retweeted persona** will be

Tweet 3 I just witnessed a DRIVER of a CAR run a Red Light. Time to fire off an Op-Ed to the @chicagotribune calling ALL drivers scofflaws! #bikeCHI
Tweet 7 When the servers try and tell you a joke saying you sound like an owl and you actually start crying bc you thought someone was being mean
Tweet 10 Update: just got reprimanded for getting on tinder. This is why I have a privacy screen. I need my replacement asap

Appendix F

Accounting for Alternative Motivations for Retweeting: Construction of the Different Types of Random Retweet Persona

We replicate our analysis, while accounting for different drivers and factors that may impact retweeting decisions. We do so by using different types of random retweeted persona vectors. Below we elaborate on the construction of the different types of random retweeted personas, and specifically the three random retweeted personas based on tie strength.

Tie Strength (1): Taking into Account Only Link Characteristics

When creating each core user *u*'s $RandomReTweet_u$ document, instead of randomly sampling tweets from the self-tweets and retweets of the user's followings, we employ a stratified sampling technique.

We first define four types of possible retweets based on the types of links between the core user and the user who originally wrote the retweeted tweet:

(1) Strong tie: The core user follows the user who wrote the retweeted tweet, and that user follows the core user.

(2) Weak tie: The core user follows the user who wrote the retweeted tweet, but that user does not follow the core user.

(3) Reverse weak tie: The core user does not follow the user who wrote the retweeted tweet, but that user follows the core user.

(4) Complete weak tie: The core user does not follow the user who wrote the retweeted tweet, and that user does not follow the core user.

Note that options (3) and (4) are indeed possible options. A user does not have to directly follow another user to have the latter user's tweet appear in his home timeline. For example, if user *a* follows user *b* and user *b* follows user *c*, if user *b* retweets a tweet of user *c*, this tweet will appear in user *a*'s home timeline even if *a* does not follow *c*. Optimally, we would want to know that the tweet arrived at user *a*'s timeline via *b*. However, the retweeting route of a tweet is not information the REST API provides. Given a tweet retweeted by an core user, the REST API provides us only with the user who wrote the original tweet. For this reason we consider tie strength types (3) and (4).

Then, for each core user we compute the percentage of retweets he posts from each of the four groups. Finally, we construct the random retweeted persona, for each core user by randomly selecting potential retweets from the user's followings in a manner that maintains the same proportions across the four tie strength groups. For example, if the core user retweeted 10 tweets from strong ties, 20 tweets from weak ties, and so forth, when sampling potential retweets from the user's followings, we will randomly sample 10 tweets from the group of tweets coming from strong ties and 20 tweets from the group of tweets coming from weak ties.

Tie Strength (2): Taking into Account Link and Interaction Characteristics

We first define five types of possible retweets based on the types of links and interactions (replies and mentions) between the core user and the user who originally wrote the retweeted tweet. Specifically, this specification divides group (1) above (strong ties) into two groups, thus eventually dividing the users' retweets into **five groups** prior to conducting the stratified sampling for the construction of the random retweeted persona:

1. **Strong tie with no interaction**: The core user follows the user who wrote the retweeted tweet, and that user follows the core user. However, there is no personal interaction between them. That is, neither user mentions the other or replies to his or her tweets using the corresponding Twitter handle.

2. **Strong tie with interactions**: The core user follows the user who wrote the retweeted tweet, and that user follows the core user, and there is at least one personal interaction between the two users (reply or mention), directed either from the core user to the following or from the following to the core user.

Then, for each core user, we compute the percentage of retweets he posts from each of the five groups. Finally, we construct the random retweeted persona, for each core user, by randomly selecting potential retweets from the user's followings in a manner that maintains the same proportions across the five tie strength groups.

Tie Strength (3): Taking into Account Link and Interaction Characteristics

We first define seven types of possible retweets based on the types of links and interactions (replies and mentions) between the core user and the user who originally wrote the retweeted tweet. Specifically, this specification divides group (1) above (strong ties) into four groups, thus eventually dividing the users' retweets into **seven groups** prior to conducting the stratified sampling for the construction of the random retweeted persona:

1. **Strong tie with no interaction**: The core user follows the user who wrote the retweeted tweet, and that user follows the core user. However, there is no personal interaction between them. That is, neither user mentions the other or replies to his or her tweets using the corresponding Twitter handle.

2. **Strong tie with one sided interaction**: The core user follows the user who wrote the retweeted tweet, and that user follows the core user. There is at least one interaction initiated by the core user toward the following but no interaction initiated by the following toward the core user.

3. **Strong tie with reverse one-sided interaction**: The core user follows the user who wrote the retweeted tweet, and that user follows the core user. There is at least one interaction initiated by the following toward the core user and no interaction initiated by the core user toward the following.

4. **Strong tie with two-sided interaction**: The core user follows the user who wrote the retweeted tweet, and that user follows the core user. There is at least one interaction initiated by the core user toward the following and at least one interaction initiated by the following toward the core user.

Then, for each core user we compute the percentage of retweets he posts from each of the seven groups. Finally, we construct the random retweeted persona, for each core user, by randomly selecting potential retweets from the user's followings in a manner that maintains the same proportions across the seven tie strength groups.

Appendix G

Determining the Number of Topics in the Corpus Prior to Running LDA

As mentioned, LDA needs to be given, *a priori*, a parameter that tells it the number of topics in the corpus. Selecting a number that is too small could cause unnecessary generalizations, whereas choosing an overly large number could cause redundancy. As explained above, in this paper we use a data-driven approach to find the optimal number of topics. Since there are several metrics that have been developed to find a "good" number, and no one dominating method, we have used four different methods, all leading to similar results. In what follows, we will briefly outline the methods used and discuss the results of each method. Modeling and computations are executed using R's ldatuning package.[1]]

Method #1 (based on Griffiths and Steyvers 2004): This method is based on evaluating the model by approximating its log-likelihood (as there are clearly too many alternatives to fully estimate the log-likelihood). The suggestion of this method is to use samples from the Gibbs sampling iterations. The number of topics is then chosen to be that with the **maximum** log-likelihood approximation.

Method #2 (based on Cao et al. 2009): This method suggests a metric to select the number of topics based on the distances among different topics in the model. The method is based on the assumption that LDA performs best when the average cosine distance of topics reaches the **minimum**.

[1] See https://cran.rproject.org/web/packages/ldatuning/index.html.

Method #3 (based on Arun et al. 2010): This method is based on Symmetric K-L divergence and on the assumption that LDA can be viewed as a matrix factorization mechanism. In the proposed metric, divergence values are higher for nonoptimal numbers of topics. Thus, the optimal number of topics would be the one that yields the **minimum** score.

Method #4 (based on Deveaud et al. 2014): This approach focuses on the goal of deriving topics that differ from one another. To this end, this approach derives the number of topics based on the information divergence (using Jensen-Shannon divergence) between all pairs of topics in a given model. The model with the **maximum** divergence is said to be the best model.

Finding the Number of Topics for the Core Users' Corpus

We ran LDA on our core user corpus using Gibbs sampling with T ranging from 5 to 150, alpha = T/50, beta = 0.1 and 1000 iterations (as suggested by Griffiths and Steyvers 2004). For each LDA, run we calculated each of the four metrics. The results are normalized and presented graphically in Figure G1 which portrays each metric as a function of the number of topics. Note that for method #1 (Griffiths and Steyvers 2004) and method #4 (Deveaud et al. 2014) we looked for a maximum, whereas for method #2 (Cao et al. 2009) and method #3 (Arun et al. 2010) we focused on finding the minimum.

According to the four metrics it seems that the optimal number of topics for our corpus ranges between 15 and 35 topics[2] (getting a range is expected given that the different approaches make different assumptions regarding what a good set of topics corresponds to). As these metrics point to a range and not a single number, in what follows we present another layer of analysis that was aimed at selecting **one single optimal number of topics** to be presented in the main results. For robustness purposes, we rerun the main analysis of the paper (H1 and H2) with 15, 20, 25, 30, and 35 topics, with similar results.

Finding a Single Optimal Number of Topics by Aggregating the Different Methods

To identify a good number of topics, we aggregated the scores of the four methods for each number of topics. For the two methods in which a higher score corresponded to a better result (Deveaud et al. 2014 and Griffiths and Steyvers 2004) we took (1-score). Thus, the optimal number of topics was the one yielding the overall **minimum** score. The results of the aggregated scores are presented in Figure G2. As shown, the number of topics with the **overall minimum** score was **25**. Thus, we chose 25 to be the number of topics used for the analysis of the core users throughout the paper.

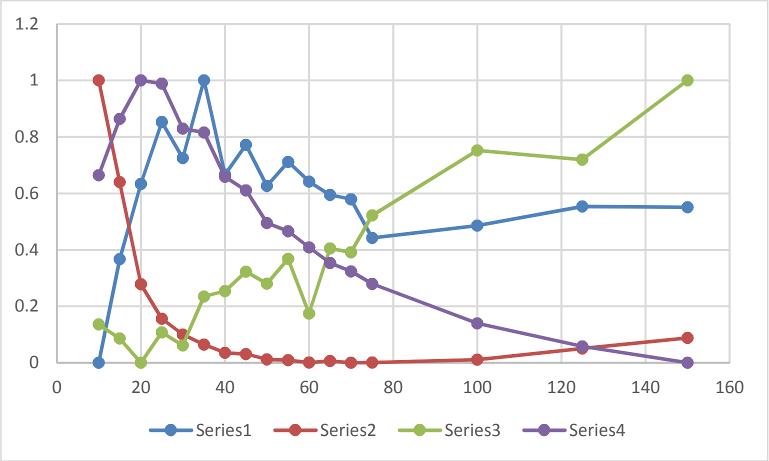


Figure G1. Measure by Number of Topics

[2] Note that Cao et al. (2009) are lower at a higher number of topics, but the delta in reduction seems to become negligible after 35 topics.

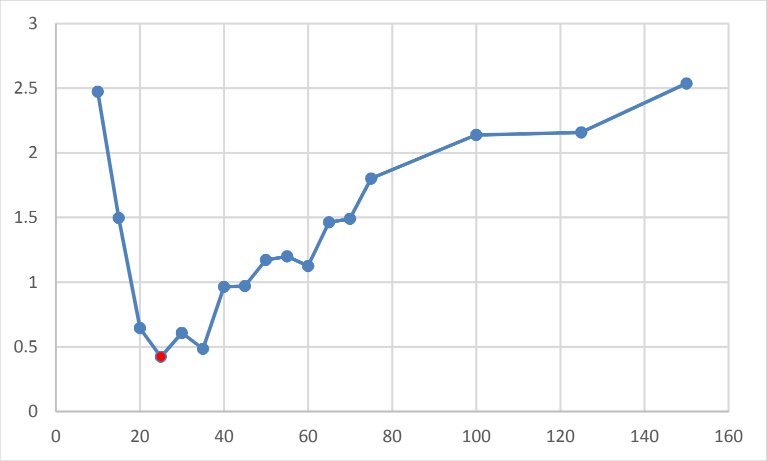


Figure G2. Sensitivity Analysis

Finding the Number of Topics for the Combined Corpus Comprising Core Users and Experts

To find the number of topics for the combined corpus we reran the analysis described above, this time on the combined data set. Figure G3 and Figure G4 present the optimal number of topics for each method and the aggregated score. As shown, the number of topics with the **overall minimum** score was **30**. Thus, we chose 30 to be the number of topics used for the analysis of the combined data set.

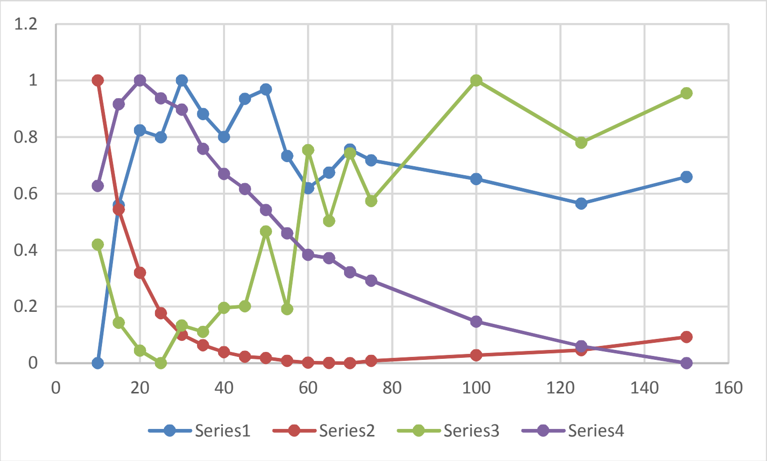


Figure G3. Measure by Number of Topics

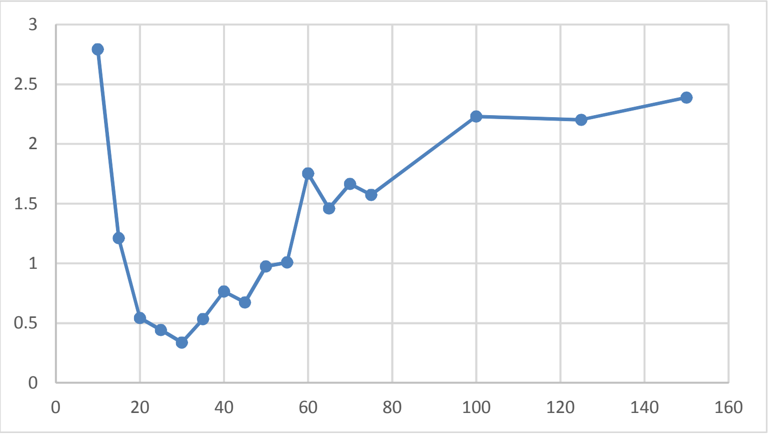


Figure G4. Sensitivity Analysis

Appendix H

Top Keywords of Topics

Table H1 presents the top 20 keywords corresponding to each of the 25 topics from the LDA run on the core users. Table H2 presents the top 20 keywords corresponding to each of the 30 topics from the LDA run on the combined data set of core users and expert users.

Table H1. Top 20 Keywords per Topic from the LDA Run with 25 Topics on the Core User Data Set

Topic 0	Social: College student life	free tonight parti night homecom black lit week colleg will atlanta fridai go ticket weekend saturdai tomorrow ladi uwg empir
Topic 1	Politics: Democrat, Bernie Sanders, police violence and social justice	polic women report nodapl kill berni peopl sander join prison war climat chang health right polici justic berniesand million syria
Topic 2	Sports: Wrestling	wwe raw match love ufc live fight sdlive wrestl tonight yr show titl win survivorseri will team hiac goldberg champion
Topic 3	Mustic: live performances and concerts	vote favorit ticket song love ama plai music artist album tonight countri rt listen show dai nowplai year tour dnce
Topic 4	Politics: presidential election	trump hillari clinton vote will elect donald realdonaldtrump obama peopl debat presid support hillaryclinton america american campaign email win fbi
Topic 5	Social: General social interactions	dai fuck love todai time work lol go good peopl feel gonna make night back watch friend hate happi shit
Topic 6	Politics: Race in politics	black trump peopl white fuck year man vote rt women obama presid girl donald will kill time call woman twitter
Topic 7	Sports: College football	happi birthdai dai school colleg great year miss todai game tomorrow senior footbal hope class week best texa tonight team
Topic 8	Social: with use of profanity	nigga shit fuck bitch lol ain ass back man wanna gotta girl love good peopl time make talk real feel

Table H1. Top 20 Keywords per Topic from the LDA Run with 25 Topics on the Core User Data Set (Continued)

Topic 9	Social: media and entertainment	will time make dai todai live good year thing great work peopl world watch love show help week read start
Topic 10	Sports: Basketball, NBA	game win plai team will season year nba back time player good watch nfl best fan man week warrior lebron
Topic 11	Entertainment: TV & movies	love watch episod season show movi book tonight film star cast comic debat happi charact read hei premier favorit fan
Topic 12	Entertainment: Travel and food	love travel happi quot beauti wine good food morn great sashaeat recip etsi vegan coffe check dai chocol yelp kitm
Topic 13	Music, rap	listen music video drop lil album song drake soundcloud bro np nigga happi kany shit birthdai rt lit rapper watch
Topic 14	Sports: Football & baseball	game win cub plai team will lead run fan tonight year footbal good time basebal todai season hit state seri
Topic 15	Social: feel-good messages	love girl life peopl make will dai thing ur time friend feel happi good person best back talk year care
Topic 16	College sport and social	student drink journalnew nursestakedc learn great beer school nurs earn join photo daytonsport untappd fairwindsbrew badg teacher educ commun rickcassano
Topic 17	Youtube shows/channels	youtub video cspanwj plai eddykenzofici artist stylish ivoteeddykenzo playlist part star movi ad regrannapp watch post makingamurder photo bt impastor
Topic 18	Sports: Soccer	goal game win score team plai player season final leagu soccer cup fan match hockei olymp nhl arsen usa messi
Topic 19	Weather	nascar todai race hurrican offic matthew polic honor florida will counti park car storm rain south fire tonight hurricanematthew back
Topic 20	Entertainment: video games	game plai pokemon video anim stream youtub final super episod draw gui appl art pokmon charact will theori jihad updat
Topic 21	Religion/faith	god peopl todai love check will automat unfollow life virgo person work lord jesu good make thing dai live time
Topic 22	The Young Turks (TYT): news and commentary program on YouTube	counti warn beach va virginia sibab sever lake thunderstorm citi theyoungturk ut mtvstarsbrunomar ride point storm cdt bruno brunomar salt
Topic 23	Popular culture	win rt follow chanc stream enter live giveawai pop dai originalfunko twitch sawyerfrdrx plai exclus will winner retweet check game
Topic 24	Beauty/cosmetics	rt win video follow dm happi gui fan palett makeup tweet song winner live show beauti vote todai birthdai amaz

Table H2. Top 20 Keywords per Topic from the LDA Run with 30 Topics on the Combined Data Set of Core Users and Expert Users

Topic 0	Religion, faith	god love will life peopl jesu lord prai bless live thing good heart faith make work give chang time world
Topic 1	Astrology	peopl todai check automat unfollow virgo person pisc sagittariu aquariu gemini tauru leo libra ari scorpio cancer feel work capricorn
Topic 2	Technology products & companies	appl market facebook googl peopl tech twitter app compani new year busi will iphon ceo startup work data report brexit
Topic 3	Politics: Race in politics	black peopl trump white vote women man fuck year rt presid obama woman men america kill donald polic hillari stop
Topic 4	Social: media and entertainment	dai time will todai make watch good live great love thing show year work world back week best night tonight
Topic 5	Wrestling	wwe raw match ufc love wrestl sdlive fight jihad tonight yr titl survivorseri team show win hiac goldberg good champion
Topic 6	Entertainment: video games	plai game stream live youtub twitch video gui team pokemon check overwatch go amaz fayde final song start peopl love
Topic 7	Mustic, live performances and concerts	sawyerfrdrx gai band song album show music plai listen vegan ur tonight art queer nowplai tour record tran sawyer sex
Topic 8	Youtube music videos	youtub video warn counti beach va eddykenzofici virginia nursestakedc artist plai stylish ivoteeddykenzo sever theyoungturk thunderstorm nurs pokemon cdt playlist
Topic 9	Middle east and war in Syria	russia syria war attack kill report israel isi russian brexit turkei world state polic aleppo forc syrian uk govern militari
Topic 10	Politics: movement for and against Trump	mtscore imwithh nevertrump donaldtrump hillaryclinton auditthevot vote amjoi msnbc notmypresid uniteblu realdonaldtrump lead knick gop flipitdem ff berlin rt joyannreid
Topic 11	Music: rap	video man lil drop drake nigga bro music album song fuck rt shit year lmao listen kany kid hit birthdai
Topic 12	Presidential election debates	trump clinton vote donald elect will peopl hillari debat presid campaign call women obama time support make voter gop year
Topic 13	Weather	polic todai offic join citi hurrican help school report live honor break matthew state park student counti power shoot commun
Topic 14	Social: feel-good messages	love girl peopl life make ur will thing dai time friend feel happi person best good talk back wanna year
Topic 15	Social: General social interactions	fuck dai love lol peopl work time todai go good gonna feel make shit night back hate watch friend thing
Topic 16	Social: with use of profanity	nigga shit fuck bitch ain lol ass love back man wanna gotta girl good time make feel talk ya real
Topic 17	Sports: Football	game win plai team will season good time footbal tonight year back week player lead todai nfl fan state start
Topic 18	Politics: Clinton email leaks and FBI	hillari trump clinton vote realdonaldtrump will hillaryclinton email maga fbi obama america elect wikileak american support peopl corrupt media win
Topic 19	Music: video and streaming	vote love favorit ama tonight artist song countri rt perform album video music show year happi live dnce taylor male
Topic 20	Popular culture	rt win follow chanc dm winner retweet palett giveawai enter pop makeup originalfunko exclus give set kit lip card kyli
Topic 21	Sports: Baseball	game cub win fan team basebal dodger plai goal seri indian hit worldseri year pitch season player score lead run
Topic 22	Music: live tour and performances	ticket love video tonight music show song tour fan gui follow fuck night listen happi live mix world amaz set

Table H2. Top 20 Keywords per Topic from the LDA Run with 30 Topics on the Combined Data Set of Core Users and Expert Users (Continued)

Topic 23	Teenage: politics, highschool sports, and entertainment	cspanwj citizenradio ut lake post citi photo salt facebook maryland drive utah il healthcar ricardoreport impastor rahde opengov allmet impastortv
Topic 24	Movies	movi episod season book star love watch show film comic review trailer war charact fan gameofthron sibab cast marvel write
Topic 25	Social: College students	happi birthdai dai school love year colleg miss todai tomorrow great game best hope week tonight good class night senior
Topic 26	Music: Alternative/Indie	np listen soundcloud free prod nowplai tonight da parti ft uwg music live mixtap night video feat periscop youtub spinrilla
Topic 27	Music: TV and youtube	raider drink beer raidern check photo live earn mtvstarsbrunomar untappd badg love good periscop jaymohrsport great bruno oakland level brunomar
Topic 28	Sports: car race	nascar race lap car win mesport regrannapp watch driver back thechas lead track bt texansch pit fan seahawk caution cup
Topic 29	Social: Student life	student school journalnew make entrepreneur learn media startup wearephoenix smallbiz teacher daytonsport book educ great start rickcassano lead will hoki

Appendix I

Accounting for Alternative Motivations for Retweeting: Results for the Different Types of Random Retweeted Persona

Below we present the results for H1 when accounting for the different factors influencing retweeting behavior.

Social Dynamics

Results When Accounting for Reciprocity

Table I1. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method

	Th **= 1**	***Th*** **= 2**	***Th*** **= 5**	***Th*** **= 10**
Mean of number of topics added via retweeted persona	2.53983	2.37387	2.14997	1.99336
Mean of number of topics added via random retweeted persona	5.99578	5.58962	4.59113	3.62764
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table I2. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method

	Th **= 0.8**	***Th*** **= 0.85**	***Th*** **= 0.9**	***Th*** **= 0.95**
Mean of number of topics added via retweeted persona	1.39	1.45	1.58	1.75
Mean of number of topics added via random retweeted persona	2.39	2.69	3.17	3.91
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Results When Accounting for Tie Strength (Definition 1)

Table I3. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method

	***Th* = 1**	***Th* = 2**	***Th* = 5**	***Th* = 10**
Mean of number of topics added via retweeted persona	2.44	2.29	2.07	1.88
Mean of number of topics added via random retweeted persona	5.90	5.53	4.49	3.50
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Table I4. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method

	***Th* = 0.8**	***Th* = 0.85**	***Th* = 0.9**	***Th=0.95***
Mean of number of topics added via retweeted persona	1.31	1.39	1.5	1.67
Mean of number of topics added via random retweeted persona	2.24	2.56	3.02	3.83
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Results When Accounting for Tie Strength (Definition 2)

Table I5. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method

	***Th* = 1**	***Th* = 2**	***Th* = 5**	***Th* = 10**
Mean of number of topics added via retweeted persona	2.28	2.18	2.04	1.95
Mean of number of topics added via random retweeted persona	5.48	5.18	4.33	3.50
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Table I6. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method

	***Th* = 0.8**	***Th* = 0.85**	***Th* = 0.9**	***Th* = 0.95**
Mean of number of topics added via retweeted persona	1.47	1.53	1.59	1.7
Mean of number of topics added via random retweeted persona	2.36	2.62	3.01	3.65
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Results When Accounting for Tie Strength (Definition 3)

Table I7. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method

	Th = 1	*Th* = 2	*Th* = 5	*Th* = 10
Mean of number of topics added via retweeted persona	2.36	2.23	2.08	1.95
Mean of number of topics added via random retweeted persona	5.44	5.15	4.23	3.39
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Table I8. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method

	Th = 0.8	*Th* = 0.85	*Th* = 0.9	*Th* = 0.95
Mean of number of topics added via retweeted persona	1.46	1.52	1.6	1.76
Mean of number of topics added via random retweeted persona	2.27	2.55	2.94	3.61
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Results When Accounting for Source (Author) Popularity

Table I9. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method

	Th = 1	*Th* = 2	*Th* = 5	*Th* = 10
Mean of number of topics added via retweeted persona	2.32829	2.20403	2.03885	1.92942
Mean of number of topics added via random retweeted persona	5.50593	5.25445	4.47805	3.70314
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Table I10. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method

	Th = 0.8	*Th* = 0.85	*Th* = 0.9	*Th* = 0.95
Mean of number of topics added via retweeted persona	1.37	1.44	1.53	1.66
Mean of number of topics added via random retweeted persona	2.47	2.77	3.21	3.83
Wilcoxon signed-rank test	*p* < 0.001	*p* < 0.001	*p* < 0.001	*p* < 0.001

Tweet Characteristics

Results When Accounting for Tweet Popularity

Table I11. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method

	***Th* = 1**	***Th* = 2**	***Th* = 5**	***Th* = 10**
Mean of number of topics added via retweeted persona	2.47914	2.30736	2.11012	1.94571
Mean of number of topics added via random retweeted persona	5.34785	5.14601	4.46503	3.64908
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table I12. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method

	***Th* = 0.8**	***Th* = 0.85**	***Th* = 0.9**	***Th* = 0.95**
Mean of number of topics added via retweeted persona	1.38	1.47	1.58	1.77
Mean of number of topics added via random retweeted persona	2.53	2.83	3.24	3.87
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Results When Accounting for Retweetability

Table I13. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Counts Method

	***Th* = 1**	***Th* = 2**	***Th* = 5**	***Th* = 10**
Mean of number of topics added via retweeted persona	2.37	2.23	1.99	1.81
Mean of number of topics added via random retweeted persona	5.66	5.46	4.67	3.81
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table I14. Mean Number of Topics Added via Retweets Versus Mean Number of Topics Added via Random Retweets: Percentage Method

	***Th* = 0.8**	***Th* = 0.85**	***Th* = 0.9**	***Th* = 0.95**
Mean of number of topics added via retweeted persona	1.3	1.36	1.49	1.63
Mean of number of topics added via random retweeted persona	2.67	2.98	3.39	4.03
Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Appendix J

Robustness Analysis: Distance Between Added Topics and Self-Produced Topics

For robustness purposes we examine another aspect of the similarity between users' self-produced personas and their retweeted personas. Specifically, we empirically study the similarity between topics added via retweets and those in the users' self-tweets, hypothesizing that they will be closely related to one another. In this analysis, we use topic similarity measures to show that the topics that users add via their retweets are more similar to the topics in their self-tweets than are those added via random retweets.

Specifically, we do the following: First, we create a distance table comparing all pairs of topics. Recall that the topics produced by LDA are multinomial distributions over the words in the corpus, and as such are suitable for comparison using methods that measure similarity (or dissimilarity) between distributions. For each pair of topics in our data set, we compute the distance between the two topics by calculating the *Jensen-Shannon divergence* (J-S divergence) between their corresponding distributions (that is, (25 × 24)/2 comparisons). J-S divergence is a popular measure for dissimilarity between two probability distributions (Aletras and Stevenson 2014). We use J-S divergence with the base 2 logarithm, which results in a number between 0 and 1, where 0 reflects identical probabilities, and 1 reflects orthogonal probabilities.

Second, for each user, we use two different measures to compute the distance between the topics that were added via the retweeted persona and the topics in the user's self-tweets. One of the measures is based on minimal distance and the other is based on average distance.

Third, for each user, we use the same two measures to compute the distance between the topics that were added via the random-retweeted persona and the topics in the user's self-tweets.

Finally, for each user and for each distance measure, we compare the distance obtained for actual retweets (the outcome of the second step) with the distance obtained for the random retweets (the outcome of the third step) to understand the relative dissimilarity between the topics in the user's self-tweets and the topics in her retweets.

We present the results of the comparison in Table J2 and Table J2. Complete details on the J-S divergence procedure and the two different distance measures are provided in Appendix K.

As can be seen in Tables J1 and J2, for both distance measures and all thresholds (corresponding to the counts method and the percentage method), we find that the topics added via the user's actual retweets are closer to the topics of her self-tweets than are those added via the random retweets. These findings provide further support to H1 in showing that, beyond the fact that users add few topics, the topics that a user does add are comparatively similar to his or her self-produced topics.

Table J1. Distance Between Topics in Self-Produced Persona and Topics Added via Retweets or Random Retweets: Counts Method

	Measure A				Measure B			
	Th = 1	*Th* = 2	*Th* = 5	*Th* = 10	*Th* = 1	*Th* = 2	*Th* = 5	*Th* = 10
Mean divergence of topics added by retweets to topics in self-tweets (RT-divergence)	0.54	0.5	0.44	0.38	0.36	0.33	0.29	0.26
Mean divergence of topics added by random retweets to topics in self-tweets (random-RT-divergence)	0.65	0.63	0.56	0.49	0.47	0.45	0.41	0.35
Wilcoxon between RT-divergence and random-RT-divergence	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table J2. Similarity Between Topics in Self-Produced Persona and Topics Added via Retweets or Random Retweets: Percentage Method

	Measure A				Measure B			
	Th = 80%	*Th* = 85%	*Th* = 90%	*Th* = 95%	*Th* = 80%	*Th* = 85%	*Th* = 90%	*Th* = 95%
Mean divergence of topics added by retweets to topics in self-tweets (RT-divergence)	0.29	0.31	0.35	0.40	0.19	0.20	0.22	0.25
Mean divergence of topics added by random retweets to topics in self-tweets (random-RT-divergence)	0.38	0.42	0.48	0.55	0.27	0.30	0.34	0.39
Wilcoxon between RT-divergence and random-RT-divergence	$p < 0.001$	$p <0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Appendix K

H1: Robustness: Methods Used to Determine Similarity Between Added Topics and Self-Tweeted Topics

We calculate similarity between the topics in the self-produced persona and the topics added via actual retweets or via random retweets. We find that the new topics added via users' retweets are indeed more similar to those discussed in their self-tweets than are the new topics added via random retweets.

Specifically, we do the following:

(1) We create a distance table comparing all pairs of topics. Recall that the topics produced by LDA are multinomial distributions over the words in the corpus, and as such are suitable for comparison using methods that measure similarity between distributions. We compute similarity using the J-S divergence between each pair of topics in our data set (that is, (25 × 24)/2 comparisons). J-S divergence is appropriate for comparing the LDA output vectors, as they are by definition probability vectors (that is, each vector sums to 1). We use the J-S divergence with the base 2 logarithm, which results in a number between 0 and 1, where 0 reflects identical probabilities, and 1 reflects orthogonal probabilities. We find that the pair with the maximum distance is topic 8 and topic 22, with a J-S score of 0.92. The pair with the minimal distance is topic 5 and topic 15, with a J-S score of 0.27. A histogram of the distances is presented in Figure K1. As can be seen, most topics are quite distinct.

(2) Then, for each user, we compute the distance between the topics that were added via the retweets and the topics in the user's self-tweets. In fact, this was done using two different measures: The first measure (denoted **measure A**), simply computes the average of distances between each topic in the self-tweets and each added topic. For example, if Jane tweets about topics A, B, and C and adds topics D and E, we compute the distance for Jane as the average of the distances A-D, A-E, B-D, B-E, C-D, and C-E. In the second measure (denoted **measure B**), we average the minimal distance between each added topic and the topics discussed in the self-tweets. For example, if Jane tweets about topics A, B, and C and adds topics D and E, we average the min of (D-A, D-B, and D-C) and the min of (E-A, E-B, and E-C).

(3) We compute the distance between the topics added via the random retweets and the topics in the user's self-tweets. As in (2), we use measures (A) and (B) to compute the distances between the topics added via the random retweets and the self-tweets.

The result are presented in the main text. We find that the topics added via the user's real retweets are closer to his self-tweets than are those added via the random retweets.

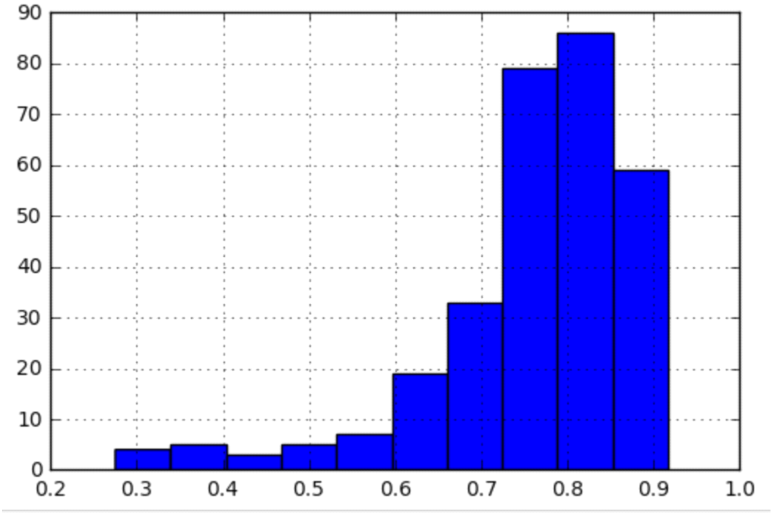


Figure K1. Histogram of Distances

References

Aletras, N., and Stevenson, M. 2014. "Measuring the Similarity between Automatically Generated Topics," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 22-27.

Arun, R., Suresh, V., Veni Madhavan, C. E., and Narasimha Murthy, M. N. 2010. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations," in *Advances in Knowledge Discovery and Data Mining*, M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi (eds.), Berlin: Springer, pp. 391-402.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. 2009. "A Density-Based Method for Adaptive LDA Model Selection," *Neurocomputing—16th European Symposium on Artificial Neural Networks 2008*, pp. 1775-1781.

Deveaud, R., SanJuan, E., and Bellot, P. 2014. "Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval," *Document numérique* (17:1), pp. 61-84.

Griffiths, T. L., and Steyvers, M. 2004. "Finding Scientific Topics," *Proceedings of the National Academy of Sciences* (101, Supplement 1), pp. 5228-5235.