Using Retweets to Shape our Online Persona: a Topic Modeling Approach

Completed Research Paper

Hilah Geva
Tel-Aviv University
Ramat Aviv, Israel
Hilahlev@mail.tau.ac.il

Gal Oestreicher-Singer
Tel-Aviv University
Ramat Aviv, Israel
galos@post.tau.ac.il

Maytal Saar-Tsechansky
University of Austin at Texas
Austin, TX 78712, United States
Maytal.Saar-Tsechansky@mccombs.utexas.edu

Abstract

Online socializing technologies have given rise to new social behaviors. We focus on the effect of reiteration tool (tools that enable us to redistribute a copy of content that others have posted) on users' online persona building. Specifically, we study retweeting behavior on Twitter and ask: (1) Do users expand the breadth of topics they discuss? (2) Do users change the distribution of the topics they discuss? (3) Does the behaviors of experts differ from those of non-experts?

We use data about 2,435 non-expert users, and 415 expert users and the users whom they follow and use LDA topic modeling to derive the topics in both self-tweets and re-tweets. We find that users rarely add new topics when retweeting and they do not alter significantly the distribution of topics. Also, this tendency is stronger among expert users, indicating that they rely more on their own words for impression management.

Keywords: Social media, Economics of information systems, Social technology

Introduction

Online social networking technologies have changed our social interactions in many important ways. In particular, these technologies have given rise to new social behaviors, with the *like* on Facebook being a prominent example. In this work, we examine how online platform users employ social networking tools to present themselves and shape their online presence. We focus on a specific type of social behavior that has emerged as a result of new social technologies: the ability to reiterate a friend's activity, that is, to redistribute an exact copy of content that he or she has posted online (e.g., words, videos, or pictures). In effect, reiteration tools such as the "Retweet" function on Twitter and the "Share" functions on Facebook and Google+ enable the user to leverage someone else's self-expression to enhance his or her own image as reflected by the online platform. While one can attempt to remember a comment or a joke and repeat it to one's friends offline, distributing a copy of the original message is effortless, and without loss of the message original's integrity. Indeed, according to Forbes.com as of January 2014, 25-30% of all Twitter activities were retweets, and this frequent use of the tool attests to its popular appeal[1].

While the technology by which individuals shape their online presence is new, the question of how individuals present themselves in different social contexts has been studied for decades. Interest in the concept of *self-presentation* can be traced to Erving Goffman's (1959) seminal work titled *The Presentation of Self in Everyday Life*. Self-presentation, or *impression management*, is defined as "the process by which individuals attempt to control the impressions others form of them" in social situations (Leary and Kowalski 1990). When engaging in impression management, individuals take into account "both the target audience and the context of the social interaction," and the process involves "making choices about what information" to divulge (Toma et al. 2008). In the marketing literature, impression management is often referred to as "personal branding" (Labrecque et al. 2011).

Since many of our social interactions and events take place online, it is important to understand how impressions are formed and managed in this environment, in which individuals have "the ability to create images of themselves for social purposes without being constrained by time or space" (Rosenberg et al. 2011). Indeed, many studies have shown that individuals are conscious of how they present themselves online and engage in strategic behavior when doing so (Rosenberg et al. 2011; Labrecque et al. 2011; Counts and Stecher 2009; Brivio and Ibarra 2009; Toma et al. 2008). Rosenberg et al. (2011) have claimed that Facebook users adopt different self-presentation tactics, depending on their personality traits. For instance, they suggest that "Facebook users who have a strong desire to be liked by others… [are] likely to use role-modeling tactics". In the context of online dating, Toma et al. (2008) have shown that individuals lie about height and weight in a strategic and intentional manner.

The capacity to engage in impression management online is enhanced by the availability of **technology enabled tools**, enabling more complex and richer opportunities for impression management. Image manipulation, the ability to publicize one's relationship status, the "like" on Facebook, and the retweet on Twitter are only a few of such popular features introduced into our social lives by online social networks. These tools also allow individuals to manage their impressions in new ways and offer individuals a rich set of means by which to shape the image they wish to project.

In this work we focus on the role that reiteration tools play in the process of impression management, specifically in the context of the Twitter social network. The distinguishing property of the reiteration tool is that it enables users to complement their own, self-produced content—their so-called *self-produced persona*—with content created by others. In other words, this tool allows the user to take content reflecting the personality and interests of others and make it her own. If we assume that the topics discussed by a user in her self-produced tweets (referred to herein as "self-tweets") are a proxy for the spectrum of topics that the user is able to self-produce, then when the user augments her own content by retweeting content produced by others—thereby creating her so-called *full persona*—she is faced with a choice: On the one hand, the user may use the reiteration tool to increase the number of topics about which she converses (i.e., the *breadth* of her online persona), so as to produce a full persona that is more diverse than the persona reflected in her self-tweets. On the other hand, she might use this tool to add content relating to topics she already addresses, thereby increasing the *depth* of her online persona. Herein, we seek to empirically explore the choices that Twitter users tend to make when retweeting content.

Prior research suggests that either type of behavior may be plausible. Specifically, some recent work suggests that users seek to create a rich online persona by providing varied and diverse content, a capacity that is enhanced by the ease of disseminating the words of others. For example, Shi et al. (2014) show that users are more likely to retweet tweets originating from weak ties, because they may carry newer information compared with tweets produced by stronger ties. They explain this behavior on the basis of the proposition that reputation-enhancement is a motivation for retweeting activities, and that the novelty of information is key to users' valuation. Peng et al. (2014) find similar evidence in the context of Digg.com. In our context, this may mean that individuals may assume the words of others in an attempt to enhance their voice by delivering content on new topics they do not produce content on themselves.

[1] See http://www.forbes.com/sites/jaymcgregor/2014/03/31/retweets-are-up-replies-are-down-how-twitter-has-evolved-in-the-last-seven-years/#4f9d34ab7b62.

On the other hand, marketing literature on personal branding suggests that although the digital age fosters the freedom to explore multiple personas, it is advantageous that one's "*personal branding message be clear and consistent, creating an air of authenticity*". According to this literature, in both traditional (Holt 2004) and personal branding (Kaputa 2005), authenticity enhances message receptivity (Labrecque et al., 2011). Moreover, it has been suggested that while it might be technically easy to create multiple personas in the online world, the public nature of the information in networks such as Twitter and Facebook makes it increasingly difficult and costly to successfully manage multiple online personas (Labrecque et al. 2011; Back et al. 2010). In our context, this may mean that users are more likely to use the reiterating tool to enhance their self-produced persona by retweeting additional content about topics they produce content on themselves. Such behavior may serve to enrich the user's persona while also keeping it focused and consistent.

To examine how users utilize the retweet function when engaging in impression management on Twitter, we investigate three research questions. Our **first research question** asks: Do users utilize retweets on Twitter to expand the breadth of topics they broadcast about, by adding topics not discussed in their self-tweets, or do they use it to enhance the depth of their self-produced persona, adding further content relating to the topics they already discuss in their self-tweets?

Even without expanding the breadth of topics in his repertoire, a user may enrich his persona by shifting the focus of his discussion, namely, by increasing the volume of the content he broadcasts in some topics more than in other topics. In light of this proposition, we also study the role of retweets in the distribution of topics the user addresses. In particular, our **second research question** asks: Do users utilize the retweet option to change the distribution of the topics they discuss in their self-tweets? Specifically, are the topics most frequently discussed by the user's self-produced persona (i.e., the topics addressed in his self-tweets) remain the same topics that are most frequently discussed by the user's full persona, comprising his self-tweets augmented by his retweets?

The behaviors we discussed above may differ across different types of users. Hence, **our third research question** focuses on users' heterogeneity by asking whether the behaviors we analyze differ between expert users (specifically, bloggers on prominent blogging sites, who are likely to engage extensively and purposefully in online impression management) and other, non-expert users.

We analyzed data taken from Twitter over a period of 3 weeks in 2015, with regard to 2,435 *non-expert core* users (defined as users who tweet regularly and are likely to use the platform for personal rather than professional purposes) and the users whom they followed. For each user, we compiled the user's self-tweets, as well as his or her retweets. We applied Latent Dirichlet Allocation (LDA) topic modeling (Blei et al. 2003) to derive the topics that users discussed in their self-produced personas (self-tweets) and in their full personas (self-tweets and retweets combined). This enabled us to capture at the individual user level both the variety of topics being discussed and the relative volume of discussion corresponding to each persona.

Given that users do not randomly choose the people whom they follow, and given that the topics that the Twitter population discusses are not randomly and evenly distributed, we must account for the possibility that the scope of topics to which users are exposed, and thus the topics that are available for them to retweet, might be biased. One potential source of such bias is the well-documented homophily phenomenon, indicating that users are likely to follow others who are similar to themselves, and who tweet about topics that they themselves discuss. To address this issue, for each individual, we constructed a *potential full persona* (also referred to as a *random full persona*): a persona constructed on the basis of the user's self-tweets, combined with a set of retweets that are randomly drawn from the user's incoming tweets; the number of retweets in the latter set is equal to the number of retweets that the user actually broadcasted.

Analysis of this data set suggests that users rarely add new topics to their profile when retweeting; instead, they enrich their persona by deepening the discussion of topics that they address in their self-tweets. Furthermore, we find that a user's retweets do remarkably little to alter the distribution of topics she discusses in her self-tweets.

Finally, we repeat our analysis and compare the behavior of non-expert core users' to that of more experienced users in the domain—expert core users (bloggers on prominent blogging sites, as noted above). We find that both non-expert and expert users tend to add few new topics via retweets, and that this tendency is stronger among expert users. Specifically, on average, experts retweet less often and add fewer topics per tweet, indicating that they rely more on their own words when engaging in impression management.

A detailed literature review is available upon request.

Data Collections and Language Processing

For the purpose of this study we collected publically available data from Twitter. We selected Twitter because of its' wide popularity, the public availability of data, and the inherent emphasis on retweeting behavior. IS researchers have been increasingly emphasizing the importance of studying Twitter data. For example, Mousavi and Gu (2014), show that Twitter activity influences politicians' law-making decisions; Oh et al. (2013), study citizen-driven information-processing through Twitter services during social crises; and Hill et al. (2013), focus on the use of the Twitter network for brand and TV audience predictions. Most relevant to our context is the work of Shi et al. (2014), who show that users are more likely to retweet tweets originating from weaker ties, because they may carry newer information, which may be perceived by followers as more valuable and thus may better serve to enhance users' reputation.

In this section, we first describe the data collection process, and then outline the language processing procedure used to isolate the topics that users discussed.

Data Collection

We collected a snapshot of 2435 non-expert twitter users, hereinafter referred to as *non-expert core users* (or *NE-core users*) and the users whom they follow, hereinafter referred to as *followings*, by querying twitter's REST API[2]. For each user we collected the user timeline, including up to the 3200 of the user's most recent self-tweets and re-tweets[3].

We selected the NE-core users according to the following three-step process, using Twitter's streaming API[4]: First, we deployed a listener to collect roughly 2 million tweets produced by US users[5] during a 24-hour window in December 2014. Second, we randomly drew a sample of 20,000 tweets that originated from 17,794 unique authors. Finally, of these potential NE-core users, we randomly selected 3,000 users, after excluding inactive users[6] and users with exceptionally high or low (top or bottom 15%[7]) number of followers or followings. This was done in an effort to eliminate users who use Twitter in a limited capacity or as a promotional platform. We subsequently eliminated 565 additional users from the set of NE-core users, due to changes in their privacy settings during the data collection window, or due to an insufficient number of English tweets in their profiles.

Then, during a 3-week window in September 2015, we used Twitter's REST API to collect information about those 2,435 NE-core users and each of their followings (over 1 million users in total). Specifically, for each NE-core user, we collected the *user timeline*—the tweets and retweets posted by the user—as well as the user's number of followings and number of followers. Note that Twitter's REST API provides the *user timeline* but not the *home timeline*—the set of tweets displayed for the user to view. Therefore, in order to reconstruct each NE-core user's home timeline (that is, to know which tweets a NE-core user was able to view), we collected the NE-core user's own timeline as well as the timelines of each of her followings. We reconstruct the *home timeline* of a user by merging all the tweets posted by the core user's followings into one timeline, sorting the tweets by creation date and removing duplicates. Consequently, our data set included not only the self-tweets and retweets of each NE-core user, but also the tweets that

[2] See https://dev.twitter.com/rest/public

[3] We note that the 3200 limitation is imposed by the Twitter REST API. We traced Tweets up to 6 month old, stopping once we reach the 3200 limit. As a result, for some of the less active users or newer twitter users there may be fewer than 3200 Tweets.

[4] See https://dev.twitter.com/streaming/overview

[5] This is based on the user's self-reported location.

[6] We define a user as inactive if his *user timeline* meets one of the followings criteria: It has (1) fewer than 200 self-tweets (2) fewer than 200 re-tweets (3) fewer than 15 retweets or fewer than 15 self-tweets in each of the three months prior to the date of collection (This criterion was not used for users who had more than 2800 tweets in total).

[7] Determining the distribution of number of followings and followers on twitter was done using a random sample of 39,162 users. The 15th and 85th percentiles used to identify users with "exceptionally high or low number of followings or followers" were taken from this distribution.

each NE-core user received but did not retweet. Additionally we collected each following's number of followings and followers.

In addition to collecting these data, we constructed a second data set (consisting of the same type of data) for 415 *expert* Twitter users, hereinafter referred to as *expert core users* (or *E-core users*). These data were collected during a 10-day window in October 2015. The process of selecting E-core users was motivated by the assumption that bloggers on prominent blog sites, who also have Twitter accounts, are likely to be so-called "expert" Twitter user. Therefore, we first chose 12 blogging websites. Second, we manually identified lists from the websites' Twitter pages that contained the twitter accounts of the bloggers. For example *Huffington Post's* Twitter account created many lists, among them a "Tech-politics bloggers" list. The full list of blogs names and the description of the lists was removed due to space limitations and is available upon request.

Using Twitter's REST API, we then collected the user profiles of the members listed in each of the lists we identified. Among these users, we selected only those who met the following two conditions: 1) the user's Twitter profile description contained at least one of the following words: *write*, *report*, *blog*, *journal*, *editor*, *column*, *correspondent* or *tweets*; assuming those words indicate being an actual expert bloggers. 2) the number of followers (users who follow them) was larger than 1000. This process resulted in the selection of 530 E-core users.

For each E-core user, we collected the same data collected for NE-core users, namely, the user's *user timeline*, the user's followings' *user timelines,* the number of followings, and the number of followers (of both the E-core users and their followings). As with the NE-core users, some E-core users were filtered out in the data collection process due to privacy settings or due to an insufficient number of English tweets in their profiles. Our final data set comprised 415 E-core users. See Table 1 for descriptive statistics of our data.

Table 1. Descriptive statistics

	Followings count (Mean) (Median)	Followers count (Mean) (Median)	Followings of Followings count (Mean) (Median)	Followers of Followings count (Mean) (Median)	Self-tweet count (Mean) (Median)	Retweet count (Mean) (Median)
Non-expert core users	518 465	663 593	2167 414	19497 556	437.13 358	419.52 316
Expert core users	1419 1129	40070 5232	2951 689	45110 1771	585.12 436	304.43 173

Table 1.

Language Processing

Our empirical analysis builds on an examination of the number of topics talked about and the topic distributions in users' self-produced persona, full persona, and random full persona. In this section we discuss the language processing performed on our collected data. Full details on the formal operationalization of each type of persona are provided in the Empirical Methodology section below.

We used LDA (Latent Dirichlet Allocation) topic modeling (Blei et al. 2003) to derive topics arising from users' sets of tweets. Rather than propose a set of topics to use in the analysis of tweets and user personas, LDA learns topics from a data corpus. That is, the only input (except the documents) is the number of topics to be learned. Specifically, LDA topic modeling models each text document in a corpus as a mixture of an underlying set of topics. Given a corpus of text documents, LDA infers both the topics exhibited in the corpus as well as the distribution (mixture) of topics in each document. The topics exhibited in the corpus are represented as a multinomial distribution over words in the corpus. In this paper we use an implementation of topic modeling by Andrew McCallum's Mallet software[8].

[8] http://mallet.cs.umass.edu/

Before running the LDA on the texts of the tweets, we applied the Porter stemmer and removed content that did not carry meaning, including stop words, URLs, Twitter-style URLs (URLs that begin with "http://t.co/"), words composed of only digits, non-alphanumeric characters, single characters, and non-Latin characters. Further, we excluded tweets that contained fewer than three words.

Prior work on topic modeling for microblogs (e.g., Mehrotra et al. 2013) has established that the inherent short length of tweets undermines LDA learning of coherent topics, and that grouping of tweets (such as by user) prior to applying LDA facilitates topic learning. Because we study the relationship between self-tweeting and retweeting for an individual user, we group tweets by author and by type (self-tweet versus retweet). Therefore, for each core user, either an expert or non-expert, we created the following text documents:

- *SelfTweet$_u$* – containing the texts of all user *u's* self-tweets that were written in English[9].
- *ReTweet$_u$* – containing the texts of all user *u*'s retweets that were written in English.

After creating the *SelfTweet* and *ReTweet* documents, we filtered out core users for whom either of the documents was smaller than 1000 bytes. This was done to facilitate a feasible run of the LDA.

For each user, we also created a document to serve as a baseline, against which the user's retweeting behavior could be compared. As discussed above, the purpose of the baseline is to control for bias in the content of the tweets to which users are exposed. Thus, each baseline document is user-specific and includes only tweets the user has received, that is, tweets that were generated by Twitter users that the core user is following.

The goal of this user specific baseline is to enable us to control for the network homophily in our analysis. Accordingly, for each user we constructed the following additional document:

- *RandomReTweet$_u$* – containing a random sample of tweets drawn from user *u*'s followings. These tweets are assumed to be representative of the tweets user *u* could have potentially re-tweeted had she re-tweeted randomly from the tweets of her followings. The number of tweets included in this document matches the number of tweets in user *u*'s *ReTweet$_u$* document.

This process resulted in 7,305 documents for the 2,435 NE-core users, and 1,245 documents for the 415 E-core users.

We then performed two separate LDA runs (with 20 topics each run) – one on all the documents of the NE-core users (7,305 documents) and one on all the documents the E-core users (1,245 documents). For robustness purposes, we performed several addition analyses: a) we reconstructed the main analysis of the paper using a 50-topic LDA run. Results of the 50-topics LDA run are reported in appendix A. Results were significant and similar in direction to those obtained in the main analysis (i.e. the analysis presented in the remainder of the paper which uses the 20-topic LDA run). b) We performed a third LDA run on a combined data set of expert and non-expert core users. The results of this analysis were in the same directions as those conducted separately on each group.[10].

Each LDA run produces a set of topics and a topic distribution vector for each of the documents, where each element in a given vector captures the portion of the content in the corresponding document attributed to a specific topic. It should be noted that for each LDA run the topic distribution vectors attributed to each user's self-tweet, retweet and random retweet documents refer to the same set of topics. We labeled the vectors produced by the LDA as follows:

- *SelfTweet_vector$_u$* –topic distribution of user *u's SelfTweet$_u$ document.*
- *ReTweet_vector$_u$* - topic distribution of user *u's ReTweet$_u$ document.*
- *RandomReTweet_vector$_u$* - topic distribution of user *u's RandomReTweet$_u$ document.*

[9] To identify English-language tweets, we rely on Twitter's own assignment of tweets to languages and filter out any tweet marked by the platform as non-English.

[10] Our analysis in this paper is conducted on the results of LDA runs on textual data from two sets of users: NE-core users and E-core users. For robustness purposes we repeated the data processing, language processing and main analyses conducted in RQ1 and RQ2 on a combined set of NE-core users and E-core users that consists of the 415 E-core users and 415 NE-core users who were chosen from the pool of 2,435 NE-core users. Results of RQ1 and RQ2 analyses conducted on this combined group of users are in the same direction as the results of the analysis conducted separately on each data set.

Because tweets are of similar length, these vectors can also be interpreted as an estimation of the percentage of tweets in the document that correspond to each topic. Given that we are interested in capturing and comparing the number of topics discussed by users, it is useful to convert the topic probability distributions into tweet *counts* per topic. Specifically, an estimate of the number of tweets in which a given topic is discussed by a user can be captured by the product of the frequency of the topic within the core user's tweets and the number of tweets by that user. We therefore constructed and computed the following vectors:

- *SelfTweetCount_vector*$_u$ is the *SelfTweet_vector*$_u$ times the number of self-tweets posted by user *u*[11].
- *ReTweetCount_vector*$_u$ is the *ReTweet_vector*$_u$ times the number of retweets posted by user *u*.
- *RandomReTweetCount_vector*$_u$ is the *RandomReTweet_vector*$_u$ times the number of retweets posted by user *u*.

Empirical Methodology and Results

In this section we describe our empirical methodology and findings with regard to our three research questions. Each sub-section will concentrate on one research question: First, we discuss whether NE-core users employ the retweet tool to add breadth or depth to their self-produced personas (RQ1). We then move on to examine how retweeting affects the distribution of topics these individuals discuss on Twitter (RQ2). We conclude by comparing the behavior of NE-core users and E-core users (RQ3).

RQ1 - How do Users Utilize Re-tweets? Do They Focus on Breadth or Depth?

In this section we present our analysis for RQ1, studying whether NE-core users utilize their retweets to add breadth or depth to their online personas.

We sought to analyze how each user's *full online persona* (i.e., his tweets and retweets combined) compared with his *self-produced persona* (i.e., his self-tweets). However, as discussed above, the distribution of topics to which users are exposed may be biased (e.g., because of homophily). To address this issue, we compared each user's full persona to his or her random full persona, which provides a proxy for the potential topics a user could have discussed, had he generated the same number of retweets but chosen them randomly from his incoming tweets. We note that the potential full persona uses a random set of potential retweets and hence does not represent the maximum potential number of topics per that user. Hence, when making a retweeting decision, the user can choose to add more or less topics by selecting those from his followings. Finally, we also note that if users were randomly following other users, we could have replaced the retweets with a random set of tweets from the entire Twittersphere (see Appendix B for our analysis of the level of homophily in the network).

Altogether, we compared the following persona representations for each user *u*:

- User *u*'s *self-produced persona*, expressed via the user's self-tweets. This persona is represented by the *SelfTweetCount_vector*$_u$.
- User *u*'s *full (observed) persona,* expressed via the user's self-tweets and retweets combined. This persona is represented by the vector sum of *SelfTweetCount_vector*$_u$ and *ReTweetCount_vector*$_u$ (denoted *FullObservedPersonaCount_vector*$_u$).
- User *u*'s *random full persona*, a baseline that reflects the user's hypothetical Twitter activity if the user were to retweet tweets drawn at random from his followings. This persona is represented by the vector sum of *SelfTweetCount_vector*$_u$ and *RandomReTweetCount_vector*$_u$ (denoted *RandomFullPersonaCount_vector*$_u$).

To examine the level to which users use the retweeting tool to broaden their persona, we assessed the average number of "new" topics included in the user's full observed persona that are not included in the user's self-produced persona, and compared it to the average number of topics included in a user's random full persona that are not included in the user's self-produced persona. For the purposes of this comparison, we considered a topic to be **meaningfully discussed** by a user (i.e., eligible to be included

[11] Recall that when creating the user tweet documents, we filtered out tweets that were not in English and tweets with fewer than three words. We count the number of tweets corresponding to each user after this filtering and not before.

in her persona) if the number of tweets discussing the topic exceeded a certain threshold *Th*. In the results reported below, we consider different threshold values of 1, 2, 5 and 10.

Table 2 presents the average number of topics (for different values of the threshold *Th*) included in the user's full observed persona that are not included in the user's self-produced persona (Row A) and the average number of topics included in a user's random full persona that are not included in the user's self-produced persona (Row B). Significance was evaluated via a Wilcoxon signed-rank test (Row C). For comparison, Rows D-H specify the average number of topics discussed in the users' self-produced persona (D), retweets (E), full persona (F), random retweets (G) and random full persona (H).

Table 2 shows that, on average, users discuss 5.59 to 8.42 topics in their self-produced tweets and 5.6 to 8.99 topics in their retweets, adding only 1.8 to 2 new topics when building their full persona using the retweeting tool. That is, the topics in the retweets and self-tweets frequently overlap. Moreover, we find that drawing retweets from one's followings indiscriminately (thus creating the random full persona) yields a significantly higher number of new topics that do not overlap with the topics discussed in the self-produced persona (between 3.25 and 4.7 topics on average).

Furthermore, we looked at the percentage of retweets (random retweets) that corresponded to the newly added topics in the full persona (random full persona), and we find that, on average (for the four different thresholds), this percentage moves between 7.53% and 15.83% (15.78% and 24.09%). This shows that most retweets contribute to topics already being discussed in the self-produced persona.

Taken together, these results provide evidence that, when using retweets to enrich their online personas, users tend to choose to deepen the topics they self-produce rather than expand the breadth of topics. Notably, comparison of the full observed persona to the benchmark of the random full persona suggests that this tendency reflects a choice made by users rather than limited availability of topics to retweet about. These observations support the personal branding theories that suggest that, when branding themselves, individuals wish and try to be consistent, as a means of conveying authenticity.

Table 2. Mean Number of New Topics in the Full Observed Persona (that Were not in the Self-produced Persona) versus Mean Number of New Topics in the Random Full Persona

		Threshold			
		Th= 1 tweet	*Th* = 2 tweets	*Th* = 5 tweets	*Th* = 10 tweets
A	Mean (SD) number of topics added in the full observed persona.	2.0 (1.38)	1.96 (1.36)	1.87 (1.4)	1.81 (1.43)
B	Mean (SD) number of topics **added** in the random full persona	4.7 (2.04)	4.52 (2.14)	3.84 (2.19)	3.25 (2.13)
C	Wilcoxon signed-rank test on A and B	Z =-41.27, p<0.00	Z = -40.93, p<0.00	Z = -38.44, p<0.00	Z =-34.78, p<0.00
D	Mean (SD) number of topics in self-produces tweets	8.42 (1.8)	7.66 (1.83)	6.59 (1.9)	5.59 (1.99)
E	Mean (SD) number of topics in retweets	8.99 (1.96)	8.16 (2.08)	6.87 (2.3)	5.6 (2.47)
F	Mean (SD) number of topics in full persona	10.42 (1.81)	9.62 (1.87)	8.46 (2.0)	7.41 (2.13)
G	Mean (SD) number of topics in random retweets	12.23 (2.46)	11.22 (2.76)	9.22 (3.21)	7.28 (3.43)
H	Mean (SD) number of topics in random full persona	13.13 (2.11)	12.18 (2.33)	10.43 (2.58)	8.84 (2.66)

Table 2.

For each threshold value, we also computed the number of NE-core users for whom the number of topics added by random retweets from followings was larger than that added by the user's actual retweets. The purpose of this analysis was to ensure that the average difference in new topics added per user cannot be attributed to a few atypical users. The results are shown in Figure 1. For each threshold *Th* the blue bar in Figure 1 shows the number of NE-core users for whom the number of new topics introduced by random retweets is larger than that introduced by the users' actual retweets, by at least *D*, with *D* ranging between 1 and 3. Considering the difference in the number of new topics introduced provides further insight into the differences between the behaviors of users' actual full personas compared with their random full personas. Specifically, for each threshold value, in more than 68% of instances the random full persona adds at least one more new topic than does the actual full persona, and in more than 40% of cases the random full persona adds at least two more new topics than does the full persona. Further, we find that for the threshold values of 1, 2 and 5, the random full persona adds at least 3 additional new topics in more than 36% of instances. These results further strengthen our conclusion that the majority of users do not use their retweets to greatly change or enrich the variety of topics they write about; rather they utilize the tool in such a way that enables them to portray a consistent and cohesive full persona.

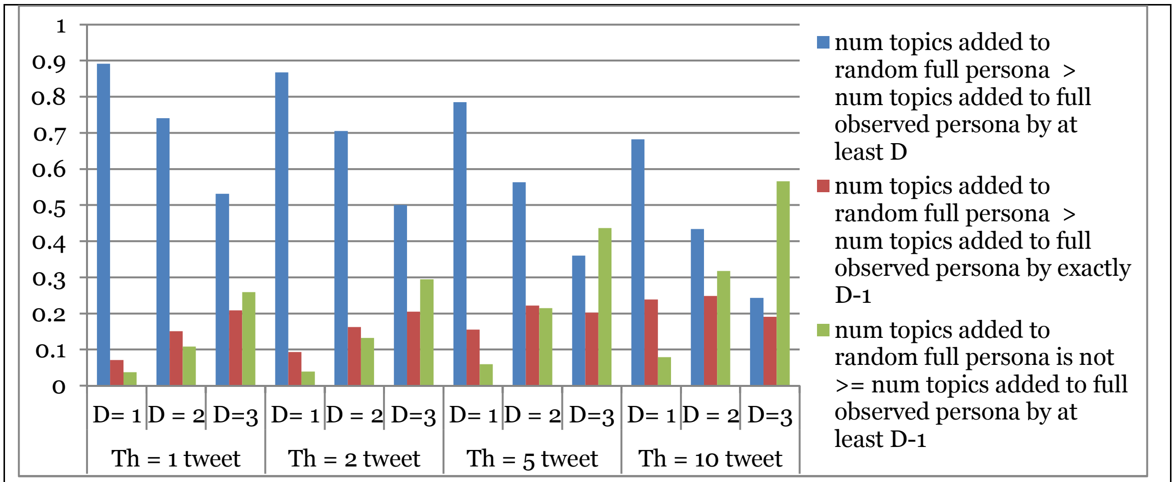

Figure 1. Full Observed Persona versus Random Full Persona

Robustness Analysis for RQ1

To support the validity of our results, we ran robustness analyses on our data. The results of these analyses are similar in both direction and magnitude to the results presented thus far. In what follows we present a brief summary of these analyses.

Controlling for Content Bias: Our first robustness analysis controls for content bias, taking into account the possibility that some tweets may be inherently less likely to be retweeted, and that enabling these tweets to be included in the random full persona might bias the results. Such tweets might include messages with very specific personal content (e.g., "it's my birthday!"). To control for this bias, we ran a second analysis in which the random retweets (i.e., the tweets grouped in *RandomReTweet$_u$*) were selected only from the followings' retweets (as opposed to the followings' retweets and self-tweets). The underlying assumption is that as a whole the type of content portrayed in re-tweets (tweets that have been found to be worthy of retweeting by at least one user) appeals to a more general public and might be different from the content posted in self tweets, which might be more personal. In this analysis, the procedures used for data processing, language processing (including a new LDA run on the newly-constructed documents) and statistical analysis were the same as those described above. The results were similar in both direction and magnitude to the results reported for RQ1. The mean number of topics added to the user's self-produced persona by the user's actual re-tweets is smaller than the number of topics added when tweets are drawn randomly from one's followings re-tweets (1.9-2.02 new topics as opposed to 3.85-4.99 new topics, respectively).

Controlling for Tie Strength: It is possible that when shaping their online personas, users decide what to retweet not on the basis of content but rather on the basis of who posted the content. For example, it may be that users simply retweet content coming from close friends (strong ties)[12]. Such a

[12] Note, that as mentioned in the literature review section, Shi et al. (2014) show that users actually tweet more from weak ties. Either way, we should control for tie strength in our analysis.

tendency might exacerbate the potential homophily bias discussed above—it is reasonable to expect that closer friends tweet about more similar content as compared with users who share weak ties. If this is the case, similarities between the topics discussed in a user's self-tweets and retweets may simply be an artifact of retweeting content that comes from close friends.

To rule out the tie strength explanation, we re-ran the analyses conducted for RQ1, while controlling for tie strength between each NE-core user and the users he retweets. To this end, we drew the random retweets (which are grouped in *RandomReTweet*$_u$) in a way that maintains the proportions across possible types of tie strength found in the actual retweets. For example, if a user retweeted 10 tweets from strong ties and 20 tweets from weak ties, when sampling the random retweets we sampled 10 random retweets from strong ties and 20 random retweets from weak ties. For a more detailed description of the results and the mechanism by which we controlled for tie strength, see Appendix C. Results were once again similar in both direction and magnitude to the results reported above.

RQ2 - Testing for Attention Redistribution

In the previous section we showed that users' full observed personas resemble their self-produced personas in terms of the topics discussed. In this section, we investigate whether the distribution of tweets across different topics differs between the two personas. Indeed, enriching one's self-produced persona may also take the form of changing the distribution of content across the same set of topics, e.g., by making the distribution more even, or by emphasizing certain topics at the expense of others. For example, one could imagine a situation in which the tweets a user generates by herself focus exclusively on 5 topics – a,b,c,d,e – where the two most discussed topics are **a** and **b**. When examining the user's full persona (both her self-tweets and retweets), we can consider two possibilities: either the relative proportions of the topics remain roughly the same (where **a** and b remain the most discussed topics), or the distribution of content across the topics may change entirely, such that the prominence of topics a and b decreases.

We used two empirical methods to address this question: First, we looked at the overlap between the topics most frequently discussed in users' self-produced personas, as compared with their full personas. Second, we compared the full topic-distribution vectors (obtained from the LDA) corresponding to the two types of personas.

A. Overlap of Topics Most Frequently Discussed

We measured the overlap between the top *Y* topics discussed in each user's self-produced persona (specifically, in his *SelfTweetCount_vector*) and the top *Y* topics discussed in that user's full observed persona (specifically, in his *FullObservedPersonaCount_vector*), where *Y* is an integer ranging between 3 and 7. We note that we count the number of overlapping topics in the set, regardless of the inter position of the topics. For example, for Y=3, if topics a, b and c are ranked 1, 2 and 3, respectively, in the self-produced persona, and they are ranked 3, 1 and 2 in the full persona, this will be considered a full overlap.

The results are presented in Table 3. We observe that for all values of Y, the median and mean of the number of overlapping topics are very close to *Y*, indicating that retweets do not elevate less-discussed topics to the top of the list. For example, when we examine the top 3 topics in the self-produced persona, we find that on average 2.42 (median 3) of these topics are also among the top 3 topics of the full observed persona. That

Table 3. Mean and Median of Number of Overlapping Topics Discussed Most Frequently in Self-produced and Full Observed Persona

	Number of top topics = 3	Number of top topics = 4	Number of top topics = 5	Number of top topics = 6	Number of top topics = 7
Mean	2.42	3.36	4.29	5.17	6.07
Median	3	3	4	5	6

Table 3

is, over 50% of the users did not change any of the top 3 discussed topics. Similarly, an examination of the top 6 topics in the self-produced persona shows that, on average, 5.17 (median 5) of the topics are also among the top 6 topics of the full observed persona.

B. Comparison of the Full Distribution of Topics

Our second approach takes into account the distribution of the entire set of topics that each user discusses, and not just the topics discussed most frequently. Specifically, we compare users' different personas using the distribution vectors obtained from the LDA.

As our analysis thus far only used count vectors—that is, vectors that represent the number of tweets in each topic—we must first create distribution vectors. To this end, we simply divide the count vectors for each user's full persona (or random full persona) by the user's total number of self-tweets and retweets. Specifically, we define the following new distribution vectors:

- $FullObservedPersona\_vector_u = \frac{FullObservedPersonaCount\_vector_u}{SelfTweet\_count_u + ReTweet\_count_u}$
- $RandomFullObservedPersona\_vector_u = \frac{RandomFullPersonaCount\_vector_u}{SelfTweet\_count_u + RandomReTweet\_count_u}$

Next, we compute the extent to which each user's self-produced tweets (represented by the *SelfTweet_vector*, which is a distribution vector attained from the LDA) are similar to the tweets of his full persona (*FullObservedPersona_vector*), in terms of the distribution of the topics addressed. Finally, we examine this similarity measure against a benchmark similarity measure, namely, the similarity between the distribution of the topics discussed in the user's self-produced tweets and the distribution of the topics discussed by his random full persona (*RandomFullObservedPersona_vector*).

We compute similarity using the *Jensen-Shannon Divergence* (JS-Divergence), which is a popular measure for similarity between two probability distributions. JS-Divergence has been used previously to compare differences between LDA probability distributions (Aletras and Stevenson 2014). It 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 JS-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.

For each user *u* we measure JS-Divergence for the two pairs of personas:

- *SelfTweet_vector and FullObservedPersona_vector*
- *SelfTweet_vector and RandomFullObservedPersona_vector*.

We find that the average JS-Divergence level between users' self-tweets and their full persona is 0.058, indicating that in absolute terms, users' self-produced personas are highly similar to their full personas, in terms of the overall distribution of topics. We further find that the JS-Divergence between users' self-tweets and their random full persona is 0.072, which is significantly higher (Significance was evaluated via a Wilcoxon signed-rank test; z = -24.535, p<0.00), indicating lower similarity. These results suggest that users retweet in a manner that produces a consistent persona, and, as indicated by the comparison with the benchmark (random full persona), this tendency is not an artifact brought on by a lack of possibility (represented by the random full persona).

These results strengthen our general conclusion from our analysis of RQ1 that, when retweeting, users tend to maintain personas that are consistent with their self-produced personas.

RQ3 - Heterogeneity Analysis: Comparing the Retweeting Behavior of Expert and Non-Expert Users

Our discussion and finding thus far have focused on non-expert users. In this section we expend our analysis to a somewhat different type of users: expect twitter users (more on how those are defined in what follows). Specifically, in this section we focus on user heterogeneity by (a) repeating the analysis conducted for RQ1 and RQ2 using a data set collected from expert (E-core) users, and (b) comparing the results of the analyses with those corresponding to non-expert (NE-core) users.

A. Expert Users' Behavior Analysis

The results of the RQ1 analysis for E-core users are reported in Table 4, and are similar in their direction to the results presented for NE-core users: We find that, on average and for different values of *Th*, E-core users introduced 0.87 to 1.28 new topics through their retweeting activity. Further, we find that drawing retweets randomly from users' followings yields a significantly higher number of new topics (between 2.18 and 5.26 on average).

Table 4. Mean Number of New Topics in the Full Observed Persona (that Were not in the Self-produced Persona) versus Mean Number of New Topics in the Random Full Persona

	Column A	Column B	
Threshold	Mean (SD) of number of topics added in full observed persona	Mean (SD) of number of topics added in random full persona	Wilcoxon signed-rank test
Th = 1 tweet	1.28 (1.29)	5.26 (2.77)	Z= 17.53, P<0.00
Th = 2 tweets	1.09 (1.26)	4.72 (3.01)	Z= 17.19, P<0.00
Th = 5 tweets	0.89 (1.16)	3.27 (2.84)	Z = 15.66, P<0.00
Th = 10 tweets	0.87 (1.11)	2.18 (2.36)	Z = 12.56, P<0.00

Table 4

The results of the RQ2 analysis for E-core users are reported in Table 5. We find that the expert users, like the non-experts, do not use their retweets to change the relative distribution of topics in their self-produced personas. In fact, looking at the medians in the analysis of topics most frequently discussed, we see that at least half of the E-core users maintain the same top topics in their full personas and in their self-produced persona. Analysis of the full distribution of topics also yields results in the same direction as the results obtained for the non-experts: We find that the average JS-Divergence value between the topic distributions of E-core users' self-tweets and of their full personas is 0.01, whereas the average JS-Divergence value between the topic distributions of their self-tweets and of their random full persona is 0.032 (Significance was evaluated via a Wilcoxon signed-rank test; Z= -17.362, p<0.00). Further, we find that for 96% of E-core users, the distribution of the self-tweets is more similar to the distribution of the full persona than to the distribution of the random full persona.

Table 5. Mean and Median of Number of Overlapping Topics Discussed Most Frequently in Expert Users' Self-produced and Full Observed Personas

	Number of top topics = 3	Number of top topics = 4	Number of top topics = 5	Number of top topics = 6	Number of top topics = 7
Mean	2.87	3.76	4.68	5.57	6.42
Median	3	4	5	6	7

Table 5

We can therefore conclude that, on average, expert users, like non-expert users, maintain full personas that are very similar to and consistent with their self-produced personas.

B. Comparing Expert and Non-Expert Users

When comparing the results of the RQ1 analysis for the two types of users, it seems that while both experts and non-experts add few topics to their self-produced personas, expert users are more extreme in this respect. Specifically, whereas non-expert users add between 1.81 and 2 new topics, expert users add only 0.87 to 1.28 new topics.

However, this comparison, which is based on the groups' tweet-count vectors, is subject to certain shortcomings. Recall that the construction of the count vectors of each user relies on the number of self-tweets and retweets the user posted. That is, if one has more tweets there can be more topics, and similarly if one has more retweets there is more potential for adding topics. Accordingly, we can only compare the results of the two groups if we believe the numbers of self-tweets and retweets posted by experts are similar to the corresponding numbers of tweets posted by non-experts. This, however, does not seem to be the case. The average numbers of self-tweets and retweets of non-experts are 437.13 and 419.52, respectively, whereas the average numbers of self-tweets and retweets of experts are 585.12 and 304.43, respectively[13]. Unfortunately, we have no way of controlling for this bias in our analyses. Hence, the comparison reported above should be interpreted with caution.

On the other hand, comparison of the results of the analyses corresponding to RQ2 does not suffer from such bias, as the vectors being compared reflect distributions of topics rather than absolute numbers of tweets.

When comparing the users' self-produced personas with their full personas in terms of overlap between the topics most frequently discussed, we see that while neither group's retweets do much to change the topic distribution, expert users are more extreme in the extent to which they focus on a consistent set of topics. For example, if we look at the top 3 discussed topics, we see that on average, for experts, 2.87 topics overlap between the self-produced persona and the full persona, as compared with 2.4 topics for non-experts.

Regarding the analysis of the full distribution of topics, we can see that, on average, the similarity between expert users' self-produced personas and their full personas (average JS-Divergence is 0.01) is greater than the corresponding similarity for non-expert users (average JS-Divergence is 0.058). Next, for both experts and non-experts, we compute the ratio between (i) the JS-Divergence of the self-produced persona and full persona; and (ii) the JS-Divergence of the self-produced persona and the random full persona. We find that this ratio is 1.24 for the non-experts and 3.2 for the experts. This means that, on average, compared with non-experts, experts show greater similarity between their self-produced personas and their full personas, both in absolute terms and also when controlling for the users possibility to change their persona.

To sum up, we find that both expert and non-expert users tend to present consistent personas on Twitter, meaning that their retweets add few topics to their self-produced personas, and do little to alter the distribution of topics addressed in users' self-tweets. However, our comparisons, coupled with our observation that expert users use the retweet option less frequently compared with non-expert users, lead us to conclude that this tendency is stronger among expert users, indicating that they rely more on their own words in their personal branding.

Conclusion

Retweeting is one of many examples of new technology-enabled social interactions in which users can reiterate the content of other users and make it their own. This paper focuses on how individuals utilize the retweeting tool when constructing their online persona on the social network Twitter.

Our first research question asked whether non-expert individuals use this reiteration tool to enrich the spectrum of topics they converse about in terms of breadth—adding more topics to their self-produced persona—or in terms of depth—adding more content to topics that they discuss in their own tweets. Our second research question examined how retweeting affects the distribution of topics addressed by non-expert individuals i.e., we asked whether these individuals use the retweets tool to alter the relative amount of tweets across topics. Finally, our third research question compared the impression management behavior of expert and non-expert Twitter users.

We analyzed data from 2,435 non-expert Twitter users and their followings, as well as from 415 expert users and their followings. Results of the analysis for RQ1 show that in the course of constructing their full personas, users tend to use the retweet option to enrich their self-produced personas in terms of depth, retweeting more about topics that they themselves already discuss. Our analysis for RQ2 indicates that users rarely use the retweet option to alter the distribution of tweets across the topics they discuss in their self-produced personas. This was found to be true both when focusing on the topics most frequently discussed, and when observing the full distribution of topics. From these findings we conclude that when presenting themselves online, users tend to use the retweeting tool to present full personas that are

[13] Recall, these are the numbers of tweets left after the data and language processing.

largely consistent with their self-produced personas. Finally, our analysis for RQ3 suggests that the behavior of expert users is similar to that of non-expert users, but that the former show a stronger tendency than the latter to maintain full personas that are consistent with their self-produced personas, relying more on their own words when managing their personal brands online.

Implications

Our work carries important theoretical and managerial implications. From a theoretical perspective, our results show that, in spite of recent work suggesting that users may seek to create colorful online personas, using varied and diverse content, in practice users choose to utilize their retweets to deepen, rather than broaden, their self-produced personas. This behavior supports the personal branding theories that suggest that when branding themselves, individuals strive to be consistent and convey an air of authenticity. More broadly, we contribute to the growing literature that aims to improve our understanding of how social interactions and behaviors are affected by the introduction of new digital technological tools in online sociotechnical ecosystems.

From a managerial perspective, we find that users tend to retweet content on topics they are familiar with (in terms of being able to produce content about those topics). This observation provides new insights into the processes of information sharing on social media networks and can inform the design of messages to facilitate the propagation of information and content[14].

Further, our analysis shows that, compared with non-experts, expert users have a stronger tendency to rely on their own words and present a consistent and unified persona on Twitter. Given that expert users can be considered as being more successful on the Twittersphere, these findings might indicate that maintaining a consistent persona may perhaps lead to a more successful Twitter experience. However, this is no more than a first indication for our data does not allow us to untangle the causality mechanism between presenting a consistent persona and being a successful Twitter user.

Limitations and Path Forward

We acknowledge that our work has certain limitations. First, our analysis in this paper was done using a snapshot of users' Twitter activity. This data set did not enable us to control for the effect of time on the evolution and construction of an online persona. Likewise, it prevents us from studying the interplay between each user's self-tweets and retweets over time. Further, a snapshot analysis restricts us from being able to untangle the causality mechanism between the way in which users present themselves on Twitter and their level of success on the network. Access to a more complete panel data of the users behavior could shed light and might allow for identification of the causality mechanism. We therefore believe that future work should focus on obtaining and analyzing a comprehensive panel data set that includes all, or most, activity of Twitter users, starting the moment they create their accounts.

Additionally, currently our study concentrates on examining a user's self-tweets and retweets in relation to the basket of tweets she receives. However, "impression management" or "personal branding" can also be explored with respect to her followers (rather than the users whom she follows). Simply stated, if my retweets were aimed at my followers, it is very likely that I take them (their interests) into account in deciding what to retweet. Future work should therefore focus on understanding the self-persona of the core user's followers and see how the user's retweeting behavior is related to the topic mix (breadth and depth) of her followers. Specifically, it should focus on whether users' retweeting behaviors vary with the interests of their audience and how the core user's choice of topics affects their likelihood to be retweeted by their followers. Finally, our study focuses on one social network. While it would be interesting to compare how users construct their online personas in similar social networks such as Facebook, we find it more compelling to study persona construction in more professional networks.

[14] For example, Sun et al. (2014) provide compelling evidence as to the importance of message design.

Appendix

Appendix A: Results for RQ1 and RQ2 with a 50-Topic LDA Run

For robustness purposes we reconstruct the analysis of RQ1 and RQ2 for our NE-core users and for our E-core users[15] using a 50-topic LDA run (instead of the 20-topic LDA run used in the main analysis). Results are in the same direction as those obtained in the main analysis, strengthening our conclusion that users tend to maintain and present a rather consistent persona.

Results for NE-core Users

The results of the RQ1 analysis for NE-core with a 50-topic LDA run are reported in Table 6, and are similar in their direction to the results presented for NE-core users with a 20-topic LDA run: We find that, on average and for different values of *Th*, NE-core users introduced 3.61 to 4.23 new topics through their retweeting activity. Further, we find that drawing retweets randomly from users' followings yields a significantly higher number of new topics (between 5.5 and 9.04 on average).

Table 6. Mean Number of New Topics in the Full Observed Persona (that Were not in the Self-produced Persona) versus Mean Number of New Topics in the Random Full Persona

	Column A	Column B	
Threshold	Mean (SD) of number of topics added in full observed persona	Mean (SD) of number of topics added in random full persona	Wilcoxon signed-rank test
Th = 1 tweet	4.23 (2.25)	9.04 (3.59)	Z=-42.135, P<0.00
Th = 2 tweets	4.08 (2.35)	8.48 (3.79)	Z=41.549, P<0.00
Th = 5 tweets	3.82 (2.46)	6.91 (3.95)	Z =38.514, P<0.00
Th = 10 tweets	3.61 (2.53)	5.5 (3.76)	Z =31.593, P<0.00

Table 6

The results of the RQ2 analysis for NE-core users with a 50-topic LDA run are reported in Table 7. We find that the NE-core do not use their retweets to change the relative distribution of topics in their self-produced personas. Analysis of the full distribution of topics also yields results in the same direction as the results obtained in the main analysis: We find that the average JS-Divergence value between the topic distributions of NE-core users' self-tweets and of their full personas is 0.068, whereas the average JS-Divergence value between the topic distributions of their self-tweets and of their random full persona is 0.082 (Significance was evaluated via a Wilcoxon signed-rank test; Z= 24.716, p<0.00).

Table 7. Mean and Median of Number of Overlapping Topics Discussed Most Frequently in Expert Users' Self-produced and Full Observed Personas

	Number of top topics = 3	Number of top topics = 4	Number of top topics = 5	Number of top topics = 6	Number of top topics = 7
Mean	2.2	3.05	3.94	4.48	5.75
Median	2	3	4	5	6

Table 7

Results for E-core Users

The results of the RQ1 analysis for E-core with a 50-topic LDA run are reported in Table 8, and are similar in their direction to the results presented for E-core users with a 20-topic LDA run.

[15] Repeating RQ1 and RQ2 with a 50-topic LDA run for the E-core users is in fact repeating RQ3.a using a 50-Topic LDA run.

Table 8. Mean Number of New Topics in the Full Observed Persona (that Were not in the Self-produced Persona) versus Mean Number of New Topics in the Random Full Persona			
	Column A	Column B	
Threshold	Mean (SD) of number of topics added in full observed persona	Mean (SD) of number of topics added in random full persona	Wilcoxon signed-rank test
Th = 1 tweet	2.69(2.33)	8.92(5.29)	Z=17.457, P<0.00
Th = 2 tweets	2.28(2.21)	7.4(5.37)	Z=17.239, P<0.00
Th = 5 tweets	1.87(1.89)	4.93(4.68)	Z =14.861, P<0.00
Th = 10 tweets	1.81(1.72)	3.54(3.6)	Z =12.373, P<0.00

Table 8

The results of the RQ2 analysis for E-core users with a 50-topic LDA run are reported in Table 9. We find that the E-core do not use their retweets to change the relative distribution of topics in their self-produced personas. Analysis of the full distribution of topics also yields results in the same direction as the results obtained in the main analysis: We find that the average JS-Divergence value between the topic distributions of E-core users' self-tweets and of their full personas is 0.014, whereas the average JS-Divergence value between the topic distributions of their self-tweets and of their random full persona is 0.039 (Significance was evaluated via a Wilcoxon signed-rank test; Z= 17.151, p<0.00).

Table 9. Mean and Median of Number of Overlapping Topics Discussed Most Frequently in Expert Users' Self-produced and Full Observed Personas					
	Number of top topics = 3	Number of top topics = 4	Number of top topics = 5	Number of top topics = 6	Number of top topics = 7
Mean	2.71	3.68	4.66	5.63	6.57
Median	3	4	5	6	7

Table 9

Appendix B: Network Homophily

In what follows we establish the presence of network homophily, by showing that our NE-core users are more similar to the users they choose to follow than they are to a random sample of users. Similarity in this context is defined as resemblance in the topics users write about in their self-tweets (i.e., similarity between the users' *SelfTweet_vector* values). For this purpose, for each NE-core user *u*, we compare between (a) the user's similarity to 40 randomly-selected users from the user's followings, and (b) the user's similarity to a group of 40 users whom she does not follow (*non-followings*). As we are comparing distributions, we compute similarity using the *Jensen-Shannon Divergence* (JS-Divergence), as we do in the analysis for RQ2. We employ JS-Divergence using the base 2 logarithm such that the resultant value varies between 0 and 1, where 0 reflects identical probabilities, and 1 reflects orthogonal probabilities. We find that over 92% of the core users are more similar to a random sample of their followings than to a random sample of non-followings (Significance was evaluated via a Wilcoxon signed-rank test; Z = 40.643, p<0.00). Furthermore, the average JS-Divergence between users and their followings is 0.38, whereas the average JS-Divergence between users and non-followings is 0.52.

To conclude, it is clear that most users are exposed to tweets by users who are more similar to themselves compared with the general population of users.

Appendix C: RQ1: Robustness 2 – Controlling for Tie Strength

We control for tie strength bias in RQ1 by repeating the analysis using a reconstructed random full persona that controls for tie strength. When creating each NE-core user *u*'s $RandomReTweet_u$ document (from which the $RandomFullPersonaCount\_vector_u$ vector is ultimately constructed), instead of randomly sampling tweets from the self-tweets and retweets of the user's followings, we employ a stratified sampling technique:

A. We first define four types of possible retweeting based on the tie strength between the NE-core user and the user who wrote the re-tweeted tweet: (1)Strong tie – the NE-core user follows the user who wrote the retweeted tweet, and that user follows the NE-core user. (2)Weak tie – the NE-core user follows the user who wrote the retweeted tweet, but that user does not follow the NE-core user. (3)Reverse weak tie– the NE-core use does not follow the user who wrote the retweeted tweet, but that user follows the NE-core user. (4)Complete weak tie – the NE core user does not follow the user who wrote the retweeted tweet, and that user does not follow the NE-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.
B. For each NE-core user we compute the percentage of retweets he posts from each of the four groups.
C. For each NE-core user, we randomly select potential retweets from the user's followings in a manner that maintains the same proportions across the tie strength groups computed in B. For example, if the NE-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.

After reconstructing the $RandomReTweet_u$ for each NE-core user, we use the same data processing, language processing and analysis procedures that we used in the main analysis of RQ1 (this includes a new LDA run on the newly constructed documents). The results are similar in both direction and magnitude to the results obtained in the main analysis. Across all threshold levels and after controlling for tie strength, the mean number of topics added to the user's self-produced persona by the user's actual re-tweets is smaller than the number of topics added when tweets are drawn randomly from one's followings (1.7-2 new topics as opposed to 3-4.8 new topics, respectively). We note that if the similarity between users' retweets and self-tweets had been attributable primarily to tie strength, we would have expected to see a large decline in the number of topics a user could potentially add if he were retweeting randomly; however, the range remains very similar to that presented in the results of RQ1.

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*, 22-27

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., and Gosling, S. D. 2010. "Facebook Profiles Reflect Actual Personality, not Self-Idealization," *Psychological Science* (21), pp. 372–374.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993–1022.

Brivio, E., and Ibarra, F. C. 2009. "Self Presentation in Blogs and Social Networks," *Studies in Health Technology and Informatics* (144), pp. 113–115.

Counts, S., and Stecher, K. 2009. "Self-Presentation of Personality During Online Profile Creation," in *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM).*

Goffman, E. 1959. *The Presentation of Self in Everyday Life*, New York, NY: Anchor Books, Doubleday.

Hill, S., Benton, A., and van den Bulte, C. 2013. "When Does Social Network-Based Prediction Work? A Large Scale Analysis of Brand and TV Audience Engagement by Twitter Users," in *Proceedings of the 34th International Conference on Information Systems (ICIS)*, Milan, Italy.

Holt, D. 2004. *How Brands Become Icons: The Principles of Cultural Branding*, Boston, MA: Harvard Business School Press.

Kaputa, C. 2005. *UR a Brand! How Smart People Brand Themselves for Business Success*, Mountain View, CA: Davies-Black Publishing.

Labrecque, L. I., Markos, E., and Milne, G. R. 2011. "Online Personal Branding: Processes, Challenges, and Implications," *Journal of Interactive Marketing* (25:1), pp. 37–50.

Leary, M. R., and Kowalski, R. M. 1990. “Impression Management: A Literature Review and Two-Component Model,” *Psychological Bulletin* (107), pp. 34-47.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. 2013. “Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling,” in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 889-892.

Mousavi, R., and Gu, B. 2014. “The Impact of Twitter on Lawmakers’ Political Orientation,” APSA 2014 Annual Meeting Paper.

Oh, O., Agrawal, M., and Raghav, Rao H. 2013. "Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises," *MIS Quarterly* (37:2), pp. 123-142.

Peng, J., Agarwal, A., Hosanagar, K., and Iyengar, R. 2014. “Towards Effective Information Diffusion on Social Media Platforms: An Analysis of Dyadic Relationships,” working paper.

Rosenberg, J., and Egbert, N. 2011. “Online Impression Management: Personality Traits and Concerns for Secondary Goals as Predictors of Self-Presentation Tactics on Facebook,” *Journal of Computer-Mediated Communication* (17:1), pp. 1–18.

Shi, Z., Rui, H., and Whinston, A. B. 2014. "Content Sharing in a Social Broadcasting Environment: Evidence from Twitter," *MIS Quarterly* (38:1), pp. 407-426

Sun, T., Viswanathan, S., and Zheleva, E. 2014. “Impact of Message Design on Online Interactions: An Empirical Investigation,” in *Proceedings of the 16th International Conference on Electronic Commerce,* p. 64.

Toma, C. L., Hancock, J. T., and Ellison, N. B. 2008. “Separating Fact from Fiction: An Examination of Deceptive Self-Presentation in Online Dating Profiles,” *Personality and Social Psychology Bulletin* (34), pp. 1023–1036.