What am I not seeing? An Interactive Approach to Social Content Discovery in Microblogs

Byungkyu Kang[1], Nava Tintarev[2], Tobias Höllerer[1], and John O'Donovan[1]

[1] Dept. of Computer Science, University of California, Santa Barbara.
bkang|holl|jod@cs.ucsb.edu
[2] Dept. of Informatics and Computing, Bournemouth University.
ntintarev@bournemouth.ac.uk

Abstract. In this paper, we focus on the informational and user experience benefits of user-driven topic exploration in microblog communities, such as Twitter, in an inspectable, controllable and personalized manner. To this end, we introduce "HopTopics" – a novel interactive tool for exploring content that is popular just beyond a user's typical information horizon in a microblog, as defined by the network of individuals that they are connected to. We present results of a user study (N=122) to evaluate HopTopics with varying complexity against a typical microblog feed in both personalized and non-personalized conditions. Results show that the HopTopics system, leveraging content from both the direct and extended network of a user, succeeds in giving users a better sense of control and transparency. Moreover, participants had a poor mental model for the degree of novel content discovered when presented with non-personalized data in the Inspectable interface.

Keywords: Communities, content discovery, explanations, interfaces, microblogs, visualization

1 Introduction

Twitter is a microblogging service where users post messages (tweet) about any topic within the 140-character limit and follow others to receive their tweets. As of Feb. 2016, Twitter has around 320M active users, and 500M tweets are sent every day. This noisy, user-generated content contains valuable information. The majority (over 85%) of trending topics are headline news or persistent news [15], and Twitter is frequently used as a news beat for journalists [5].

With large amounts of noisy, user generated content, we have no choice but to rely on automated filters to compute relevant and personalized information that are small enough to avoid cognitive overload. However, once an automated information filtering mechanism of any type is applied, there is a real risk that useful, or critical information will never reach the end user. This problem is not new: there is a sweet-spot between similarity and diversity in personalization. Smyth [22] and Herlocker [11] refer to it as a general black-box problem with recommender systems, and more recently, Pariser [19] describes it as a *filter*

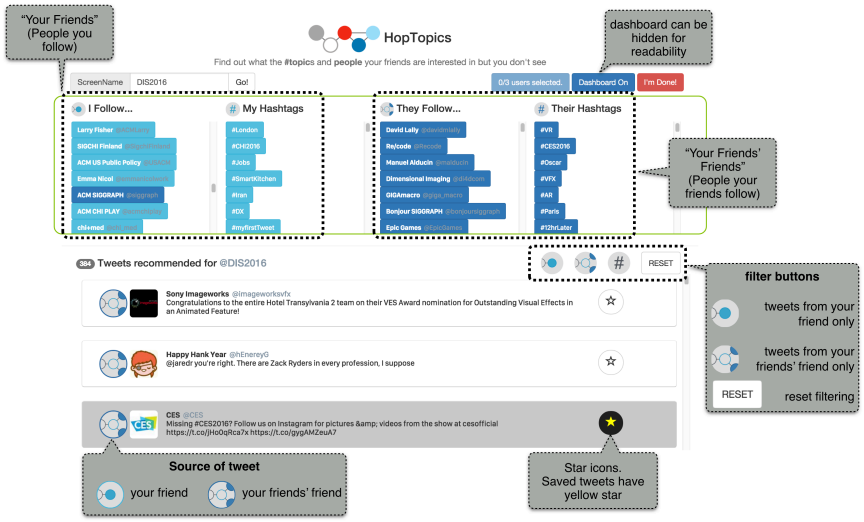


Fig. 1. Screenshot of the system (condition D), with explanatory labels. Top: Network Dashboard controlling the source of the tweets using community structure (1-hop and 2-hop followers) and topics (hashtags). Bottom: Content Viewer with resulting tweets, that can be filtered and starred by users.

bubble problem, wherein personalized filtering algorithms narrow a user's window of information.

In social networks such as Twitter, a user's information feed is populated with content from the other users that they follow directly. Here, filtering can be seen as a two step process. First, the user elects to follow another user, and second, that (second) user acts as an information curator by either authoring or propagating messages. Both steps in this process are subject to failures (*c.f.*, [9],[23]). Allowing people to see how their social network influences the information they receive may help alleviate these issues. A high profile example of the filter bubble effect occurred during the 2016 "Brexit" vote in the United Kingdom. Through automatic personalization of content via algorithms, or via selected friends serving as news curators, many people incorrectly assumed that the 'vote remain' campaign would win by a large majority, and expressed shock on Twitter after the result was announced. We believe that explanation of filtering processes, and user-interaction with information filtering algorithms can help to mitigate filter-bubble effects.

To this end, the main contribution of this paper is the introduction and evaluation of a novel interface for Twitter, *HopTopics*, that addresses the filter bubble problem. HopTopics enables users to leverage their network to source novel and potentially relevant topics from both the local and extended social network. The approach can be viewed as a hybrid of strong and weak ties (c.f., [10]) for personalized information seeking. The main scientific contribution is

the use of community structure to support content discovery, while improving control and inspectability of navigation. In our study, we use an example from journalism, but we note that the system can be generalized to most information-seeking tasks, such as understanding public opinion about political events, as in the above example.

The remainder of this paper is organized as follows: First, we present a discussion of related work. Next, we introduce the HopTopics system, including the design choices for the interactive user interface (Figure 1). This is followed by a user experiment (N=122) in which we evaluate the system on real data and users from Twitter in a 4 x 2 mixed design experiment. We conclude by discussing key results and ideas for next steps.

2 Background

To frame this research in the context of related work, we look at three key areas. First, we discuss related work on *inspectability and control in intelligent systems*. Second, we focus on inspectability and control of data in *microblogs* – a topic which is central to our research and has also received much attention in recent years. Finally, we present a discussion of related work in the area of *community based content discovery*.

2.1 Inspectability in Intelligent Systems

Mechanisms for improving inspectability and control have been introduced to different classes of intelligent systems from open learner models [6, 8], to autonomous systems [7], decision support [3, 14], and recommender systems [13, 25]. These studies have found that inspectability and control can have a positive effect on user experience as well as improved mental models.

There has been a shift toward supporting more open searches and users' understanding of novel domains evolving through use [1]. This has also meant an evolution from static explanations to more dynamic forms of explanation such as interactive visualization. For example, [26] has looked at how interaction visualization can be used to improve the effectiveness and probability of item selection when users are able to explore and interrelate multiple entities – *i.e.* items bookmarked by users, recommendations and tags.

2.2 Inspectability in Microblogs

In order to better deal with the vast amounts of user-generated content in microblogs, a number of recommender systems researchers have studied user experiences through systems that provide inspectability of and control over recommendation algorithms. Due to the brevity of microblog messages, many systems provide summary of events or trending topics with detailed explanations [16]. This unique aspect of microblogs makes both inspectability and control of recommender algorithms particularly important, since they help users to more efficiently and effectively deal with fine-grained data. Schaffer *et al.* found that in addition to receiving transparent and accurate item recommendations in a microblog, users gained information about their peers, and about the underlying algorithm through interaction with a network visualization [20]. The Eddi sys-

tem [4], a Twitter dashboard that supported topic-based browsing on Twitter, and was found to be more efficient and enjoyable way to browse an update feed than the standard chronological interface.

2.3 Community-based Content Discovery

Serendipity is defined as the act of unexpectedly encountering something fortunate. In the domain of recommender systems, one definition has been the extent to which recommended items are both useful and surprising to a user [12]. This paper investigates how exploration can be supported in a way that improves serendipity, and maintains a sense of inspectability and control. The intuitions guiding the studies in this paper are based on findings in the area of social recommendations, *i.e.*, based on people's relationships in online social networks (*e.g.*, [17]) in addition to more classical recommendation algorithms.

The *first intuition* is that weak rather than strong ties are important for content discovery. This intuition is informed by the findings of the cohesive power of weak ties in social networks, and that some information producers are more influential than others in terms of bridging communities and content [10]. Results in the area of social-based explanations also suggest that mentioning which friend(s) influence a recommendation can be beneficial (*e.g.*, [21, 27]). In this case, we support exploring immediate connections or friends, as well as friends-of-friends.

The *second intuition* is that the intersection of groups may be particularly fortuitous for the discovery of new content. This is informed by exploitation of cross-domain exploration as a means for serendipitous recommendations [2]. It is in these two intuitions that the novelty of the work in this paper lies: using community structure in microblogs to aid content discovery, by supporting inspectability and control.

3 HopTopics System

In the following sections, we first describe the UI design behind HopTopics, from the initial design used in a formative study, to the final design, shown in condition D of the main study, and in Figure 1. After that, we describe the interaction design.

The system architecture was designed to support real-time network-based, topic-specific data exploration, including caching algorithms in order to prevent exceeding the given rate limit[3], and a cluster of back-end servers to increase the number of possible concurrent requests.

3.1 Formative user study

We first conducted a formative user study (N=12) to evaluate the interface and interaction design [24]. We used a layered evaluation approach [18], focusing on the decision of an adaptation and how it was applied (in contrast to which data was collected or how it was analyzed). So, to isolate some aspects of user interface and interaction design, the HopTopics interface was evaluated using

[3] `https://dev.twitter.com/rest/public/rate-limiting`, retrieved July 2016

obfuscated (*Lorem Ipsum*) data. We briefly summarize the previous results and their impact on the interface and interaction design below.

Participants in two countries interacted with the interface in semi-structured interviews. In two iterations of the same study (n=4, n=8), we found that the interface gave users a sense of control. Users asked for an active selection of communities, and a functionality for saving individual 'favorite' users. Users found the community-based exploration feature to be particularly useful (feature retained), but re-ranking of tweets less so (feature omitted).

Based on this formative study, a number of improvements have been implemented in the system presented in this paper. The current system uses a scrolling mechanism to allow users to see more data, and clearer icons to annotate whether a tweet comes from someone the user follows, or whether it is from someone two hops away. This addresses issues with previous annotations about group membership being unclear. The current interface also considers participants' requests to better integrate with the existing twitter website: it now includes a facility to favorite tweets, and includes multimedia and URLs. As a consequence, the system contains both a *Content Viewer* (for Tweets) and a *Network Dashboard* pane (for Inspectability and Control) (*c.f.*, Section 3.2).

3.2 User Interface Design

Figure 1 shows a screen shot of the training screen for the system and indicates the various components. The dark grey speech boxes illustrate the basic components of the system. The system has two core components: 1) A *Network Dashboard* shows the active user's one and two hop network along with the topics/hashtags that are prominent in them; 2) A *Content Viewer* panel shows a collection of the messages that are derived from the current set of selections made in the dashboard view.

The *Content Viewer* shows an iconized combination of tweets from the different network groups: one-hop, two-hop, and global tweets (outside the user's network). Messages from each group are shown with a source provenance icon, shown on the left side of Figure 1. Within this viewer, participants can filter messages based on group (one-hop, two-hop, global).

Due to limited screen space for most web users, an important feature of the system is the ability to retract the Network Dashboard – which is the inspectability and control mechanism – and focus only on the Content Viewer – which contains the tweets: the information they are typically interested in. In addition to the icons, a color coding scheme is applied to the Network Dashboard to indicate links between hashtags and the groups (one-hop, two-hop, global) they originate from, as shown in dark blue and cyan in the Network Dashboard of Figure 1.

Since the number of nodes increase exponentially as one traverses the Twitter network, query complexity and data relevance were primary design considera-

Table 1. Overview of conditions. Degree of personalization is within subjects, system type is between subjects.

	Baseline	Augmented Data	Inspectable	Controllable
Non-personalized	A1	B1	C1	D1
Personalized	A2	B2	C2	D2

tions[4]. Node selections were limited to three selections for each column. The selection limit is shown dynamically on the top right of the view window.

3.3 Interaction Design

A user begins by typing a query into the system. This is typically their own Twitter ID, but could also be another user whose network they are interested in exploring. The API is queried for network and content information, which are then displayed in the Dashboard and Content panes respectively. A user can mouse-over any of the user profiles listed in the columns to find out more (*e.g.*, view the profile description and picture). Users interact with the Dashboard by first selecting up to 3 people in the left column, which consists of all the people they follow (one-hop group). As a user clicks on one of the people in the first column, the third column is populated with people who they follow (two-hop group, or friends-of-friends). When the user selects someone from the two-hop group, further hashtags get shown in the "their hashtags" column furthest to the right. This also adds more tweets in the Content Viewer.

As in typical Twitter feed interfaces, users can "star" or favorite tweets in the Content Viewer. A user can also select the 'reset' button at any point in a data exploration session to return the system to its default view of the network.

4 Experiment

In this section, we describe an experiment to evaluate versions of the interface, and the effect of personalization, using real world data. The experimental toolkit was deployed as a web service and the link was made available on Amazon Mechanical Turk (AMT).

4.1 Experiment Design

The experiment used a mixed design, as shown in Table 1. The interface variant was assigned between participants and was one of: A) Baseline - standard Twitter feed only, B) Data - augmented feed including topics mentioned by friends of friends, C) Inspectable - dashboard visible but not interactive, D) Controllable - interaction with dashboard, this is the full system introduced in the previous section. These conditions were compared between rather than within participants, in order to avoid learning and ordering effects for a specific Twitter account.

[4] In an ideal scenario, the HopTopics system would be connected via a firehose, `https://dev.twitter.com/streaming/firehose` (retrieved July 2016), connection where complex queries would not pose quite as much of a constraint. However, given limited bandwidth for our real-time experimental setup, node selections had to be restricted.

In addition we compared, within participants, the effect of personalization of the data, comparing 1) a non-personalized id (always the same id: @ACMIUI) with 2) a personalized (their own Twitter ID). The condition using the data for the non-personalized ID was always shown first. While this data was retrieved live, it was not currently in progress. This is also a relatively small community on Twitter, so the dataset is relatively static, and contains many topics that may be unfamiliar to the average Twitter user.

The motivation for using such a dataset is 1) to create familiarity with the interface through training, and 2) to have a condition where we expect participants to have a consistent level of familiarity (low) with the content, as it is not personalized. This design means, for example, that participants assigned to the Augmented Data condition would always see first B1 (non-personalized) and then B2 (personalized).

In addition to the responses we collected and computed the following indirect measures: #people saved, #tweets starred, #hashtags saved, correlation between perceived novelty and #hashtags identified as novel.

4.2 Hypotheses

We hypothesized that the system will help users discover more unexpected and useful content, and lead to better subjective perceptions when the system is a) Inspectable and Controllable (compared to two benchmarks: Baseline and Data), and when the content is personalized (compared to non-personalized content). The itemize hypotheses are described in Table 2.

Participants. Participants were recruited from the US only and were required to have a Mechanical Turk acceptance rate of greater than 90% (at least 90% of their HITs are considered of good quality by other requesters). They were required to correctly answer some filler questions, and to have a minimal degree of interaction with the system (2 minutes and 1 interaction).

155 participants completed the full study, however 33 participants were excluded from analysis as they had technical issues (most likely they interacted with the system beyond Twitter's rate limits[5]. Out of the remaining 122, the distribution across the 4 versions of the system was (A=32, B=32, C=27, D=31). The lower number of participants in conditions C and D is due to a lower completion rate in these conditions. User comments suggest that this is due to the system being slow.

The majority of participants (50%) were 25-35, 24% were 18-25, 22% were 35-50, and only 4% were 50-65. Participants were balanced across genders (49% male vs 51% female). 91% of the participants reported that they used Twitter ("Sometimes" (27%), "Often" (39%), or "All of the Time" (25%)).

4.3 Materials

Tweets were retrieved live at the time that the experiment was run (Oct 2015), and used to populate the interface described in Section 3. Tweets were either retrieved for the @ACMIUI account (in the non-personalized condition), or the

[5] `https://dev.twitter.com/rest/public/rate-limiting`,retrieved July 2016

Table 2. Overview of hypotheses (H1-H7)

H1: Perceived serendipity. "Compared to your regular twitter feed, how much does this interface help you find relevant and surprising items that you did not know about yet?" H1a: Perceived serendipity will be higher in the Inspectable and Controllable conditions. H1b: Perceived serendipity will be (slightly) higher in the personalized compared to the non-personalized conditions, across types of system.
H2: Perceived familiarity. "Compared to your regular twitter feed, how helpful is this interface for finding information that is both relevant and familiar?" H2a: Perceived familiarity will be higher in the Baseline and Data conditions. (compared to Inspectable and Controllable). H2b: Perceived familiarity will be higher in the personalized compared to the non-personalized conditions, across types of system.
H3: Perceived transparency. "The interface helped me understand where these tweets came from." Perceived transparency will be higher in the Inspectable and Controllable conditions. We do not anticipate a difference in perceived transparency between the personalized and non-personalized conditions.
H4: Perceived control. "The interface helped me change the tweets that are recommended to me." Perceived control will be higher in the Inspectable and Controllable conditions We do not anticipate a difference in perceived control between the personalized and non-personalized conditions.
H5: Content discovery. H5a: Degree of content discovery (sum of people + hashtags + tweets saved) will be greater in the Controllable condition (compared to the Interactive condition). H5b: Degree of content discovery will be greater in the personalized than the non-personalized conditions.
H6: Correctness of mental model. H6a: The correlation between perceived serendipity (subjective) and content discovery (objective) will be higher for the Controllable condition (compared to the Inspectable condition). H6b: The correlation between perceived serendipity (subjective) and content discovery (objective) will be higher for the personalized compared to the non-personalized condition.
H7: Perceived diversity. Perceived diversity will be higher in the Inspectable and Controllable conditions (compared to the Baseline and Data). We do not anticipate a difference in perceived diversity between the personalized and non-personalized conditions.

user's own Twitter id (personalized condition). The default ordering of tweets and people given by the API was used (chronological order).

4.4 Procedure

The procedure contained the following steps, described in detail below: Pre-survey ⇒ Instructions ⇒ HopTopics Non-personalized ⇒ Post-survey 1 ⇒ Instructions ⇒ HopTopics personalized ⇒ Post-survey 2. Participants started the experiment with a pre-survey[6], including demographics. They were then taken to an instruction screen (see Figure 1), where they were given an open-ended task:

Imagine that you have just taken on a new role as a freelance journalist. You need to write a few pieces for a client. You can write them on any topics you find interesting and surprising. However, you need to send your boss a short summary on these topics by tomorrow! Your job is to find people and topics that help you with your task:

∗ *Save people and hashtags by clicking on them.*

∗ *Star any tweets that you would use as the basis of the articles you are going to write.*

Once they closed the introduction screen, the main interface became visible with the non-personalized content. Participants could not move on to the next screen if they had interacted with the system for less than 2 minutes. In A and B conditions only interactions with tweets could be logged, while in conditions C and D, interactions with people and hashtags could also be logged.

Participants moved forward to a post-survey[7] after they selected the "I'm Done!" button. Here they were asked about their perceptions of the system and its contents. Next, participants were taken to the personalized variant in their condition where they were asked to enter their own Twitter ID. They performed the same task a second time, and were taken to an identical post-survey for this second interaction.

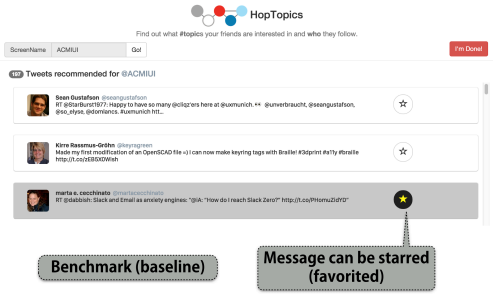
4.5 Results

The distribution of responses were not normally distributed for any of the variables, and non-parametric tests (Kruskal-Wallis or Wilcoxon) are used to consistently compare between conditions.

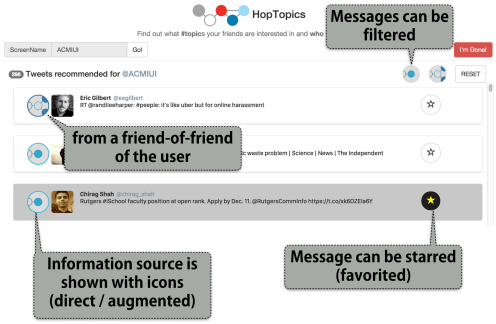
H1: Perceived serendipity. H1a. There was no significant difference between versions of the interface with regard to the degree of perceived serendipity (Kruskal-Wallis chi-squared = 5.8, df = 3, p-value = 0.12). H1b. There was also no significant difference between the perceived serendipity for the non-personalized and personalized conditions (Wilcoxon rank, W = 5876.5, p-value = 0.71). Tables 3 and 4 summarize the means. While we did not find an effect of condition on perceptions of serendipity, these differences may be revealed in a more longitudinal study.

[6] Pre-survey, `https://ucsbltsc.qualtrics.com/SE/?SID=SV_819BEUBNzqmvzed`.

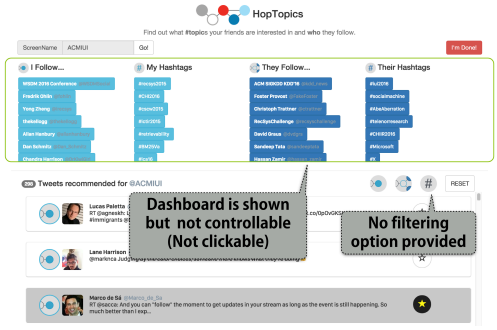
[7] Post-survey 1, `https://ucsbltsc.qualtrics.com/SE/?SID=SV_8kvPK1O6qGiW2Z7`.



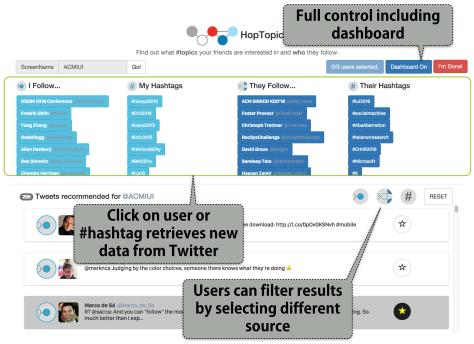
(A) Baseline: This is the typical message list view.



(B) Augmented Data: Augmented with new messages.



(C) Inspectable HopTopics with dashboard.



(D) Controllable HopTopics with interaction.

Fig. 2. Screenshots of the four between subjects system versions (A-D).

Table 3. Mean (SD) perceived serendipity by system level (A-D), (0=low, 100=high).

A	B	C	D	All
59.81 (30.58)	50.97 (32.54)	62.59 (28.95)	60.38 (26.33)	57.95 (30.34)

Table 4. Mean (SD) of perceived serendipity for personalized and non-pers. conditions.

Non-personalized	Personalized	All
57.69 (29.38)	58.22 (31.39)	57.95 (30.34)

H2: Perceived familiarity. Table 5 summarizes the mean familiarity of hashtags. H2a. In a comparison between the Inspectable and Controllable interface conditions, we did not find a significant effect on the perception of being able to find familiar tweets (Wilcoxon rank, W = 1944.5, p-value = 0.07).

Table 5. Mean (SD) of content identified as relevant and familiar. (1=low, 100=high)

	Non-Pers.	Personalized	Both
C	59.96 (28.03)	55.37 (27.69)	57.67 (27.69)
D	38.67 (34.40)	52.07 (34.18)	45.37 (34.66)
Combined	48.75 (33.05)	53.63 (31.04)	51.19 (32.02)

H2b. There was no significant difference w.r.t. perceived familiarity comparing the personalized compared to the non-personalized data (*c.f.* Table 5), across types of system (Wilcoxon, W = 1497.5, p-value = 0.47). A deeper understanding of the sets of selected hashtags, and associated user interests may be necessary to explore differences in familiarity perception.

H3: Perceived transparency. There was a significant difference between the versions of the interface (A-D) w.r.t. the degree of perceived transparency (Kruskal-Wallis chi-squared = 8.5456, df = 3, p-value < 0.05). The means in Table 6 show higher means for the Controllable and Inspectable conditions (pairwise comparisons were not significant after correction was applied). We did not anticipate a difference in perceived transparency between the personalized and non-personalized conditions.

H4: Perceived control. There was a strong and significant difference between the versions of the interface (A-D) w.r.t. the degree of perceived control (Kruskal-Wallis chi-squared = 13.562, df = 3, p-value < 0.01). Table 7 summarizes the means per condition, demonstrating a greater sense of control in the Inspectable (C) and Controllable (D) conditions. *Post-hoc* pairwise comparisons show a significant effect of conditions only between the Baseline and both the Inspectable (C) and Controllable (D) conditions (Wilcoxon, $p < 0.01$, Bonferroni corrected). We did not anticipate a difference in perceived control between the personalized and non-personalized conditions.

Table 6. Mean (SD) perceived transparency by system level (A-D), (1=low, 7=high).

A	B	C	D	All
4.45	5.12	5.41	5.45	5.09
(2.05)	(1.82)	(1.52)	(1.37)	(1.76)

Table 7. Mean (SD) perceived control by system level (A-D), (1=low, 7=high).

A	B	C	D	All
3.97	4.38	4.88	5.03	4.49
(1.67)	(1.88)	(1.63)	(1.29)	(1.71)

H5: Content discovery. Participants could not select topics or people in conditions A and B, so we compared conditions C and D only. H5a. There was a trend toward greater content discovery in the Inspectable (C) condition compared to the Controllable (D) condition (Wilcoxon rank, W = 1168.5, p-value = 0.07), but the Inspectable condition also had a much larger standard deviation. H5b. There was also no significant difference for the degree of content discovery between the non-personalized and personalized conditions (Wilcoxon rank, W = 6131.5, p-value = 0.84), means are shown in Table 9.

Table 8. Mean (SD) of content discovered by system level (C, D). (1=low, 7=high)

C	D	All
6.37	5.00	5.65
(5.17)	(3.49)	(4.40)

H6: Correctness of mental model.

Participants could not select hashtags or people in conditions A and B, so we compared conditions C and D only. H6b. There were significant correlations between perceived serendipity and degree of content discovery in the personalized condition (Table 10). This suggests that participants were better at estimating the amount of novel hashtags and tweets they were marking as favorite in this condition.

H6a (revised) *Post-hoc* comparisons show that for the Inspectable interface and the non-personalized condition this correlation was negative and significant (Spearman, $p<0.05$, rho=-0.42, Bonferroni corrected). This suggests that participants who could inspect their data, but had a non-personalized profile, were poor at estimating the amount of novel content they marked as favorites. Participants in the non-personalized condition discovered as much novel content (Table 9) as for the personalized condition, but underestimated the perceived novelty (*c.f.* Table 4).

H7: Perceived diversity. There was no significant difference between the versions of the system (A-D) w.r.t. the degree of perceived diversity (Kruskal-Wallis chi-squared = 3.8267, df = 3, p-value = 0.28). We did not anticipate a

Table 9. Mean (SD) content discovered for pers. and non-pers. conditions. (1=low, 7=high)

Non-personalized	Personalized	All
5.65 (3.81)	5.65 (4.95)	5.65 (4.40)

Table 10. Spearman correlations b/w perceived serendipity and content discovery.

Comparison	p	rho
Condition C	0.16	-0.19
Condition D	0.57	0.10
Personalized	0.02	0.15
Non-Pers.	0.16	0.08

difference in perceived diversity between the personalized and non-personalized conditions.

5 Limitations

The focus of our study was on the impact of inspectability and control of a recommender system on content discovery and user experience. The recommendation process itself was treated as a black box, enabling us to establish that the *interface and interaction* (decoupled from the algorithm) are effective in terms of user perceptions.

We did not apply our own recommendation algorithm to filter people to follow, the hashtags, or the tweets. Rather, the content was based on social network structure, and the chronological order normally used by Twitter. While Twitter is reputed to modify its ranking algorithms, details of the current tweet selection algorithm are available in their online documentation[8]. With similarity-based ranking applied to the user lists, the HopTopics system would behave similarly to a standard user-based collaborative filtering algorithm. While outwith the scope of this paper, the next natural steps will be to implement this system with different algorithms, such as user and item-based collaborative filtering, and compare user perceptions for these. Note that in our experiment, the *same* black box algorithm was used across all conditions, creating a fair comparison in a between-subjects design for the different levels of inspectability.

Our definitions of serendipity and familiarity follow the definition of *e.g.*, [12], and put an emphasis on "relevance", *e.g.*, serendipity is defined as both relevant and surprising. Given the relatively weaker personalization, through the regular Twitter content selection, and selection by social network, it is likely that there were fewer *personally relevant* items than might appear in a system that applied a better personalization algorithm. In this paper, the main point was however to compare between conditions. The results demonstrate that there were enough relevant tweets to compare (fairly) between conditions, and we were able to evaluate familiarity and serendipity. An alternative could have been

[8] `https://support.twitter.com/articles/164083`, retrieved July 2016

to ask only whether a tweet is surprising (/familiar). However, this would not capture the impact of an inspectable interface on deciding whether a tweet is relevant, instead it would answer the question if they have seen it before (or not) regardless of its informational value.

6 Conclusion and Future Work

Table 11. Significance levels, - if not significant. NA=Not Analysed.

Variable	**Interface**	**Personalization**
Serendipity	-	-
Familiarity	-	-
Transparency	0.05	NA
Control	0.01	NA
Content	-	-
Mental Model	-	0.05

In this study, we designed a novel interactive tool, called HopTopics, for social content discovery on Twitter. The system is designed to combat the filter-bubble effect by sourcing relevant information from beyond a user's typical information horizon in microblogs. Specifically, we leverage n-hop social connections and hashtags to create augmentations to the "traditional" Twitter feed that include opinions on relevant topics from both local (1 hop) and extended (>1 hop) networks. We conducted a 4 by 2 mixed design user experiment to evaluate the impact of feed augmentation, inspectability and control on a variety of UX metrics compared to a baseline, while varying personalization. The study was conducted on a crowd-sourced platform, and a fictional information gathering scenario was used to promote user engagement with the system. Our findings suggest that our interface and interaction model are significantly more effective for improving user experience in terms of both perceived transparency and perceived control, compared to baseline interfaces.

Inspectability can sometimes also be harmful. Participants in the Inspectable condition underestimated how much novel content they discovered when the twitter data was not personalized to them. While the results for the Inspectable and Controllable interface conditions were largely identical throughout the experiment, the error in mental models found for the Inspectable condition did not appear for the Controllable condition. This suggests that the Controllable interface might help users form better mental models, and that the measure of serendipitous content objectively discovered, versus the perception of serendipity, could be a good proxy for evaluating when transparency is helpful. Table 11 summarizes the key findings of the experiment. In our next experiments, we will investigate the effect of specific algorithms (*e.g.* kNN and content-based filtering) on the selection of top users and hashtags on interaction with the interface. We also plan to investigate the effect of the global/trending use of hashtags mentioned in the network on novel and relevant content discovery.

7 Acknowledgements

This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. R. A. Amar and J. T. Stasko. Knowledge precepts for design and evaluation of information visualization. *IEEE Trans. Visualization and Computer Graphics*, 11:432–442, 2005.
2. P. André, m.c. schraefel, J. Teevan, and S. T. Dumais. Discovery is never by chance: Designing for (un)serendipity. In *Creativity & Cognition*, 2009.
3. S. W. Bennett and A. C. Scott. *The Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chapter 19 - Specialized Explanations for Dosage Selection, pages 363–370. Addison-Wesley Publishing Company, 1985.
4. M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: Interactive topic-based browsing of social status streams. In *User Interface Software and Technology*, UIST '10, pages 303–312, 2010.
5. M. Broersma and T. Graham. Twitter as a news source. *Journalism Practice*, 7(4):446–464, 2013.
6. P. Brusilovsky, E. Schwarz, and G. Weber. ELM-ART: An intelligent tutoring system on world wide web. In *Intelligent Tutoring Systems*, 1996.
7. F. Cerutti, N. Tintarev, and N. Oren. Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In *ECAI*, pages 207–212, 2014.
8. V. Dimitrova. Style-olm: Interactive open learner modelling. *International Journal of Artificial Intelligence in Education*, 17(2):35–78, 2003.
9. S. Garcia Esparza, M. P. O'Mahony, and B. Smyth. Catstream: Categorising tweets for user profiling and stream filtering. In *International Conference on Intelligent User Interfaces*, IUI '13, pages 25–36, 2013.
10. M. S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.
11. J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *ACM conference on Computer supported cooperative work*, pages 241–250, 2000.
12. J. L. Herlocker, J. A. Konstan, L. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.
13. B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.
14. T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *IUI*, 2015.
15. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *International Conference on World Wide Web*, WWW '10, pages 591–600, 2010.
16. A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, 2011.

17. S. Nagulendra and J. Vassileva. Providing awareness, understanding and control of personalized stream filtering in a p2p social network. In *Conference on Collaboration and Technology (CRIWG)*, 2013.
18. A. Paramythis, S. Weibelzahl, and J. Masthoff. Layered evaluation of interactive adaptive systems: Framework and formative methods. *User Modeling and User-Adapted Interaction*, 20, 2010.
19. E. Pariser. *The filter bubble: What the Internet is hiding from you.* Penguin Books, 2011.
20. J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O'Donovan. Getting the message?: A study of explanation interfaces for microblog data analysis. In *Intelligent User Interfaces*, IUI '15, pages 345–356, 2015.
21. A. Sharma and D. Cosley. Do social explanations work? studying and modeling the effects of social explanations in recommender systems. In *World Wide Web (WWW)*, 2013.
22. B. Smyth and P. McClave. Similarity vs. diversity. In *International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development*, ICCBR '01, pages 347–361, 2001.
23. J. Teevan, M. R. Morris, and S. Azenkot. Supporting interpersonal interaction during collaborative mobile search. *Computer*, 47(3):54–57, 2014.
24. N. Tintarev, B. Kang, and J. ODonovan. Inspection mechanisms for community-based content discovery in microblogs. In *IntRS15 Joint Workshop on Interfaces and Human Decision Making for Recommender Systems at ACM Recommender Systems*, 2015.
25. N. Tintarev and J. Masthoff. *Recommender Systems Handbook (second ed.)*, chapter Explaining Recommendations: Design and Evaluation. 2015.
26. K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In *International Conference on Intelligent User Interfaces*, IUI '13, pages 351–362, 2013.
27. B. Wang, M. Ester, J. Bu, and D. Cai. Who also likes it? generating the most persuasive social explanations in recommender systems. In *AAAI Conference on Artificial Intelligence*, 2014.