RESEARCH

Open Access

CrossMark

Tag clouds with a twist: using tag clouds coloured by information's trustworthiness to support situational awareness

Jason R. C. Nurse[1*], Ioannis Agrafiotis[1], Michael Goldsmith[1], Sadie Creese[1] and Koen Lamberts[2]

* Correspondence:
jason.nurse@cs.ox.ac.uk
[1]Cyber Security Centre, Department
of Computer Science, University of
Oxford, Oxford, UK
Full list of author information is
available at the end of the article

Abstract

The amount and variety of information currently available online is astounding. Information can be found covering any subject and is accessible from any part of the globe. While this is beneficial for countless purposes, whether they be in understanding situations or for making decisions, the sheer amount of information has led to significant problems in *information overload*. As humans, we are simply unable to consume, reason about, and act on such a vast quantity of information in a timely manner. This is especially true in cases where gathering a quick understanding or awareness of a situation is desirable, or even required. In this article, therefore, we aim to investigate an approach to helping address this problem, which builds on our previous research in the area of assessing and presenting the trustworthiness of online information. Specifically, this article examines the capability of tag (or word) clouds, coloured according to the trustworthiness of the contexts in which they appear, in supporting an individual's understanding of a situation. The novelty of this work is in the application of such tag clouds to a new decision-making context, and engaging in a critical, user-based assessment of their use. To comment briefly on our findings, we note that there is potentially a significant value to be gained in the application of this technique, in providing a quick, helpful and accurate overview of a situation. This could be exploited by the public at large, but possibly even in more official investigative or crisis-management scenarios.

Keywords: Information trustworthiness, Tag clouds, Communicating trustworthiness, Situational awareness, Decision-making, Social-media content, Risk communication

Introduction

Advances in technology have reshaped the world that we live in. One major example of this is the Internet and new platforms such as Web X.0, which enable content to be accessed and published from anywhere, at any time and by anyone across the globe. Unfortunately, as a society, we have reached a point where there is simply too vast a quantity of information available for it all to be accessed and properly considered by an individual [1]. This applies to existing bodies of information but especially to information about new and ongoing events. Consider the 2013 Boston bombings where shortly after the explosions went off, tweets about the attack reached 44,000 per minute [2]. It would be impossible for an individual to read that many tweets in a timely manner, not to mention combining their information to better understand the

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 2 of 22

situation. To exacerbate matters, research and various example cases have demonstrated that information online is often of varying levels of quality and trustworthiness [3]. As a result, not only are users of online information faced with too much information to consume but they also need to be particularly careful in what information they base their decisions on.

In this paper, we aim to tackle issues that users of online information face by investigating a layered information presentation approach to support an individual's understanding of a situation. Specifically, we examine the ability of tag clouds, coloured according to the trustworthiness of the set of messages they are found in, to enhance the situational awareness of an individual. A tag cloud, or word cloud, is a technique used to visualise textual data, where word size, colour or positioning can be used to indicate characteristics of the words (e.g., frequency or prominence) in relation to the text. Considering this ability, a prime application of tag clouds has been the creation and display of summaries of large amounts of text [4]. Our approach aims to combine the summary capability of this technique with our previous work on assessing the trustworthiness of information [7, 8], so as to create specialised clouds with words (or parts thereof) coloured according to the trustworthiness of the underlying content (i.e., the different information items used to generate the cloud). Our core goal is to develop and then evaluate the use of such an approach in summarising large amounts of information and then providing individuals with a quick, helpful and accurate overview of the situation.

The remainder of this article is structured as follows. In Section 2, we reflect on existing research in the domain of information quality and trustworthiness, and visualisation techniques to communicate information. Section 3 introduces and explains the concept of coloured tag clouds and discusses how they might be used for situational awareness. We then outline the experiment that was conducted to investigate the usefulness of the coloured tag clouds in Section 4, while Section 5 presents and discusses the results of the experiment. Finally, Section 6 concludes the article and highlights avenues for future work.

Reflecting on the state of the art

The concepts of information quality and trustworthiness have attracted the interest of the research community for some time. Keeton et al. [9] focus on identifying how relevant information is for its intended use to determine its quality, whereas they perceive trustworthiness as an extension to the quality of information [10]. Wang and Strong identify various factors that can be used to determine the quality of information and how much to trust it, including accuracy (information is correct), objectivity (it is unbiased), relevancy (applicable and helpful), believability (regarded as true and credible), timeliness (age of the data is appropriate), completeness (sufficient breadth, depth and scope), and reputation (trusted in terms of source) [11]. Researchers have explored these factors further by concentrating on specific areas where trust is important. Rieh and Belkin [12] for example, focus on the quality of information on the Web and identify factors influencing people's judgement.

Defining the quality and respective trustworthiness of information is of paramount importance, but the question of how this information will be communicated to users also poses many challenges, for instance, that of *information overload*. The fields of risk communication and system usability can be used to provide solutions to some of these

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 3 of 22

issues and many researchers have proposed best practice guidance [13] and instructions on using effective visuals [14, 15]. For example, Chevalier et al. focus on visuals and more specifically, on using graphs and charts as the means to assist users in assessing the quality of content [16]. Idris et al. consider the use of traffic-lights [17], and Adler et al. advocate changing the background colour of related text to indicate content trust [18]. A more recent article by Volk et al. extend current visualisations, including some of our own, and propose a trust visualisation based on radar plots and pie charts [19]. Their approach concurrently shows multiple trust scores along with an aggregated trust score, and also includes a reliability measure for every trust score graphically, in the form of a certainty score.

To our knowledge, there is no existing research that considers the use of tag clouds as means to communicate the trustworthiness of information. Tag clouds have been heavily used in social and collaborative software [20], especially for summarising results [21]. Hearst and Rosner provide evidence that tag clouds assist in identifying what a group of people is interested in and how these interests may change over time. In addition, tag clouds have been perceived as popular and fun [22]. Other research, such as Halvey and Keane, focus on how various properties of tag clouds (i.e. alphabetization, usage of larger fonts) can be utilised to assist users in finding tags and relevant information [23]. In this article therefore, we aim to explore the various useful properties of tag clouds to assess whether they can be used to efficiently and effectively communicate the trustworthiness of online content.

Colouring of tag clouds and their context of use

Tag clouds represent an intriguing way in which textual data can be summarised and visualised; this could be especially useful where there are large volumes of data (e.g., in ongoing crisis situations). In this article, we propose combining this technique with our previous research on the topic of analysing and presenting the trustworthiness of online information [7, 8]. The presumed workflow would be as follows:

- First, information about a particular situation (for instance, an ongoing crisis) would be gathered from online sources (e.g., social media and news sites) by our existing system [8];
- Next, the system would measure the trustworthiness of each piece of information using a range of trustworthiness metrics;
- Finally, once each information item is assigned with a trustworthiness score (our previous work used a traffic-light system to indicate the score for each tweet), the items would be processed to generate a tag cloud with words coloured according to the trustworthiness of the messages where they are found.

Figure 1 presents an example of such a tag cloud based on data adapted from the 2011 UK Riots.

This specialised cloud has the same properties of a basic tag cloud, in that, the size of words indicates their frequency across the set of information items. The main difference is its use of red-, yellow- and green-coloured words to indicate the trustworthiness of the information item (e.g., Twitter post) in which they appear. In some cases, as shown in Fig. 1, where a word is used in items of varying levels of trustworthiness, we

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 4 of 22

Fig. 1 An example of a coloured tag cloud based data from the UK Riots of 2011

colour characters within the word. Therefore, green coloured characters in a word give an initial indication of the times the word is used within highly trustworthy content, and similar situations for yellow (with medium trustworthiness) and red (with low trustworthiness); this, therefore, works like a pie chart where in this case the word is the 'pie' and proportions are according to trustworthiness. Our approach to determine the number of characters to colour (green, yellow or red) is based generally on trustworthiness of the information items which contain the word and their prevalence. For instance, if most of the items that contain the word are of high trustworthiness, then the word in the tag cloud will be mostly green. If there is a roughly even split in the trustworthiness of items using the word (i.e., one third of the statements are highly trustworthy, one third are of medium trustworthiness and one third are of low trustworthiness), then the word will be coloured in the same way (i.e., one third of characters will be green, one third yellow and one third red). The only exception to this is where we have to accommodate for practical limitations, for example, where there are only a few (e.g., three) characters in a word, or only a single item representing one level of trustworthiness.

To take Fig. 1 as an example, it is clear that the most frequently mentioned word across the set of information items is 'Birmingham'. However, most of these information items are rated with low levels of trustworthiness (hence a higher percentage of red characters in the word 'Birmingham'). From this, one might infer that based on the information to hand possibly nothing is happening in that location, but instead, that there is unrest more broadly in the UK, given that content with 'UKRiot' is used mostly in medium-to-high trustworthiness content (given most of the word's characters are green or yellow). Of course, this might not truly be the case; it is important to consider the complete set of words and to delve into the detail in the related information items to form a good judgement of what might actually be occurring in such a situation. In the next section we outline an experiment which looks to assess the utility of such tag clouds and their value at assisting an individual's situational awareness.

Experimenting with coloured tag clouds

Overview and research questions

To investigate the usefulness of coloured tag clouds as a means to provide an overview of a situation, we designed a user-based experiment. There were four main activities.

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 5 of 22

The first activity involved recruiting individuals. In the second activity, we engaged with participants to brief them on coloured tag clouds, how they work (e.g., meaning of word sizing and colours) and what they aim to communicate. The next activity consisted of gathering feedback from participants on a set of pre-defined tag clouds; for this study, we used an adapted Twitter dataset from the 2011 UK Riots [24], where the tweets had been classified according to their trustworthiness (high, medium or low). In terms of feedback, we specifically asked participants to use only the information available on screen to decide what they thought was happening in the situation (represented by the tag cloud). This was then to be communicated to us in one sentence.

Participants had 4 min to consider the tag cloud and any specific information in tweets, before they needed to provide feedback. Each participant was presented with eight screens, each with a differently coloured tag cloud. After completion, participants were asked to complete a questionnaire and were interviewed to gather their general thoughts about the interface. Our research questions were:

- Can tag clouds with words coloured according to the trustworthiness of related content facilitate a quick, accurate and helpful overview of a scenario/situation?
- And, as a secondary aim, how do individuals react to these specialised tag clouds generally? Were they considered useful or were they even more confusing than facing large amounts of text?

These questions were used to guide the design and method underlying the experiment.

Participants

In total, 42 individuals (24 females, 18 males, $M_{age} = 22.57$, age range: 18–40 years) participated in the experiment. These individuals were recruited from the cities of Oxford and Coventry in the UK through the distribution of flyers and posts on notice boards. Participants were from a variety of professions and disciplines, and compensation was provided for participation.

Experiment design

The experiment was structured around eight screens, or Content sets, each displaying a coloured tag cloud. These tag clouds were generated from four sets of tweets related to specific situations of unrest during the 2011 UK Riots. The screens displayed content as follows: Content sets 1 and 2 contained tweets about unrest in Birmingham, Content sets 3 and 4 presented tweets about unrest in Tottenham, Content sets 5 and 6 focused on unrest in Manchester, and Content sets 7 and 8 covered tweets from Islington. There were 30 tweets within each set. To control the experiment, and thus ensure that participants were able to make useful conclusions about the situations presented in the coloured tag clouds, we selected specific related tweets with pre-defined levels of trustworthiness to constitute the sets. Each content set contained a majority of one level of trustworthiness (15 items) and a minority of the other two levels (ten items and five items respectively).

We set up content sets to have essentially the same information (i.e., Content sets 1 and 2, sets 3 and 4, sets 5 and 6, and sets 7 and 8) to allow us to examine whether

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 6 of 22

participants might react differently if the trustworthiness of certain information is modified (while keeping the information itself identical). For instance, does changing the asserted trustworthiness of items (and thus, the colours of words in the tag clouds) mean that participants will change their opinion on what is happening in a particular situation? A positive answer would mean that tag clouds coloured by trustworthiness might be a useful tool at accurately summarising a large amount of information.

A screenshot of the tag cloud for Content set 1 is shown in Fig. 2 – this displays the cloud on the 10" Motorola Xoom tablet used for the experiments. In these clouds, colours indicate trustworthiness and the size of a word indicates its frequency amongst the set of tweets in that set. The right side pane presents a scrollable list of the tweets used, similar to our previous experiments ([7]) where each tweet is assigned a level of trustworthiness, presented by a traffic light (green, yellow and red, relate to high, medium and low trustworthiness content respectively). Another feature of the tag cloud is that clicking/tapping words allowed filtering of the information items in the right side bar. Therefore, tapping "street" in Fig. 2 filtered the content in the list and only content with the word "street" appeared. Tapping multiple words would create a filter where items would be shown only if they contain all of the words selected. Participants were fully briefed of these options, the meaning of colours in the tag cloud and basic tag cloud properties (as mentioned above), before the experiment.

The format of the experiment involved presenting participants with the tag clouds for each of the content sets, and asking them a series of feedback questions. The first question was: "Based on the information presented on the screen, please describe in one sentence what you think is happening?" Next, participants were asked what words they would tap on to find out more about the situation. Finally, after tapping a few words and reading the filtered content in the side bar, participants were given the opportunity to revise or replace their initial sentence. All of the sentences, words, comments and feedback provided were recorded.

Fig. 2 Screenshot Content set 1's tag cloud

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 7 of 22

Results and discussion

To analyse the data gathered, we considered each content set separately and the feedback that was received from participants about it. We started by comparing the responses received for the first question. In cases where the sentences provided (or spoke about) the contents of the high and medium trusted content, this was noted as agreement with these sets. Conversely, where the sentences covered topics related to low-trustworthiness content (or indeed topics not in the data at all), this was considered as disagreement. Broadly speaking, we interpreted agreement to mean that participants were able to achieve a quick and reasonably accurate overview of the situation using the cloud. This is therefore a positive result indicating that the tag clouds were useful at supporting situational awareness.

Next, we examined the words chosen by participants from the tag cloud. We were interested in what words participants were choosing and why. For example, did participants focus on words because of their colour, size, or some other factor? We therefore tallied the words supplied and assessed the features of those words as well as the justifications supplied by individuals as to why they chose them.

Finally, a second assessment, similar to the one used with the first sentence, was applied to the second sentence. We were keen to determine whether (and how many) participants changed their sentences after actually reading content, thereby indicating that the tag cloud did not give them a good grasp of the situation. If there was a poor initial understanding of the situation followed by substantial change in sentences – particularly if new sentences were more in agreement with high and medium trustworthy content – this could indicate that the tag clouds might have been misleading. That is, instead of helping, they had led people to a potentially undesirable overview of the situation. Below we present that set-by-set analysis.

Content set 1

In Content set 1, a majority (85 %) of study participants provided statements in agreement with the high and medium trustworthy content. This was a noteworthy finding, as it supported the case for tag clouds and was evidence towards answering the research question positively. Statements that did not agree tended to focus more on the less trustworthy content, intertwined with assumptions about what could be happening in the scene. For example, in the cloud of the Content set 1 (see Fig. 2), ethnicities were mentioned, and from this, one participant deduced the summary sentence: "The riots are very violent and racism is a factor". This conclusion, however, was contrary to reports from the high and medium-trustworthiness content.

Regarding the words chosen, these are shown below – frequencies of selection are bracketed (Table 1).

Comparing the most frequently selected words to the cloud, a pattern in participants' choices was revealed. 'Birmingham', for example, was the biggest word in the cloud and was used in some highly trustworthy, but a majority of medium trustworthy content; thus likely selected for size and trustworthiness. This was confirmed once we had consulted the reasons given by participants for that word choice. Often they stated that its size played a key part, even though it was not fully trustworthy, and also in the case of 'Birmingham', the fact that it was a location. Location was important because, if

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 8 of 22

Table 1 Content set 1: Words selected by participants and their frequencies

Arrest (13)	Looting (2)
Asian (2)	Murder (30)
Better (1)	Muslim (5)
Birmingham (18)	Pakistani (4)
Breaking (2)	Police (13)
Death (3)	Riot (8)
Definitely (2)	Risk (4)
Green (1)	Street (3)
Hope (1)	UKriot (7)
Killed (4)	Violence (2)
Knife (1)	Young (1)

they were in Birmingham, they would access content containing that word to discover more about the events or potential riots in Birmingham. If they cared about another location, the word might not have been that relevant. 'Police' was also selected for its size and its mention in almost entirely trustworthy content. Additionally, participants stated that they desired to know what exactly police were doing in the situation. Were police absent, hence the riots? Were they present, or had they regained control of a riot situation?
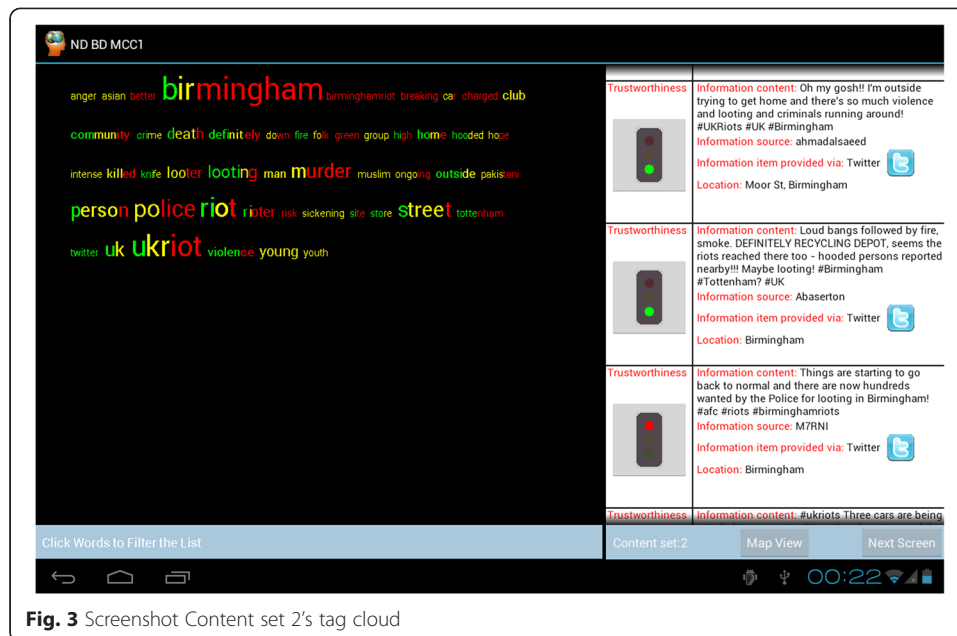
The word 'murder' was also of interest to participants. In general, the key motivational factors were: size (it was reasonably large), positive trustworthiness, and affect or emotion. According to participants, emotion was very important in their choice. They found 'murder' to be "shocking", "very serious" and highly emotive and therefore, when they saw it, they wanted to find out more. Similar reasoning was also given for words 'death' and 'killed'. Another word which was moderately popular to participants was 'arrest'. The reason reported for this was its colour – it was the only word used exclusively in highly trustworthy content. From the brief analysis of this content set, several encouraging patterns could already be seen. These emphasised the importance of factors such as size, colour, meaning and emotion, on participants' word-focus and selection decisions.

When asked about the second sentence (i.e., after they had time to tap on words and read actual content), only 24 % of participants opted to significantly change their original sentences. This suggested that a majority of individuals were satisfied that the tag cloud gave them a good initial take on the situation such that they did not need to alter their decision. Given that most of the initial sentences matched the content from trustworthy sources, this was a noteworthy finding. In situations where there was a small change, this was mainly to add more detail or further refine original opinions. As such, the cloud could still be seen to give a reasonable preliminary overview that could be suitably refined after reading the information content in detail.

Lastly, we noted that after incorporating the feedback on the second sentences, 88 % of people provided sentences that agreed with trustworthy content. This demonstrated an expected increase in understanding given the additional information supplied.

Content set 2

In Fig. 3, we present a screenshot of the system used to represent Content set 2's data. Here we found that 78 % of participants were able to supply summary sentences in

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 9 of 22

Fig. 3 Screenshot Content set 2's tag cloud

agreement with the situation's ground truth. Sentences that did not agree appeared to place emphasis on words with some level of trustworthiness but almost the same level of untrustworthiness, 'Pakistani' and 'Tottenham' are two examples. This, however, is understandable given the presence of some positive degree of trustworthiness. Other sentences focused on lower trustworthy content or made too many assumptions about what might have been occurring. Words selected are shown in Table 2.

The most prevalent three words selected for this set, were 'Birmingham', 'looting' and 'violence'. 'Birmingham' was selected for the same reasons as before, but according to participants, especially because of its size – being the biggest word – and not its trustworthiness. For 'looting', the most frequently chosen word, participants stated that its mostly green colour and slightly large size were the primary reasons for selection. We hypothesize that colour rather than size, and potentially the meaning of words were

Table 2 Content set 2: Words selected by participants and their frequencies

Anger (1)	Muslim (1)
Asian (1)	Ongoing (1)
Better (1)	Outside (2)
Birmingham (17)	Person (3)
Charged (1)	Police (5)
Community (3)	Riot (10)
Death (5)	Street (5)
Definitely (2)	Tottenham (1)
Killed (2)	Twitter (3)
Knife (4)	Ukriot (6)
Looting (20)	Violence (12)
Man (1)	Young (1)
Murder (9)	Youth (1)

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 10 of 22

also factors in participants' decisions, which is why larger words with more neutral connotations such as 'street' and 'person' did not feature as heavily.

According to participants, the preference for 'violence' was largely based on its colour and to a small extent the emotion it elicited from individuals (e.g., how violent is the situation?). This selection was interesting because it was preferred over bigger words such as 'murder'. One reason for this might be the colouring of 'murder' or more specifically, the fact that it was hardly used in content of any reasonable degree of trustworthiness. Compared to Content set 1, the impact of the difference in colouring can be clearly seen. Unsurprisingly therefore, the main reason participants gave as to why they selected 'murder' in Content set 2 was because of the emotion it elicited and on occasion, its size – not colour. The effect of colour on selection can also be seen in other words such as 'person' and 'community' (increases in trustworthiness and also selection frequency in set 2) and 'police' (decreases in trustworthiness and consequently in participants' preference).

After word selection and browsing content, only 22 % of participants made a notable change to their original sentences. Again, this was encouraging given the largely accurate initial judgements by individuals. At times, changes provided sentences in agreement with the trustworthy content whereas at other times, they disagreed. Disagreement was primarily linked to including more information than was in the content and making at times sensible but unsubstantiated assumptions. Similarly to Set 1, where smaller changes were made in the second sentences, these were for refinement purposes. Finally, we found that after being allowed to update their sentences 78 % of people were in agreement. This matched the percentage the tag cloud received on its own and could therefore be seen as a positive result – i.e., the tag cloud alone performed just as well as the cloud plus detailed content.

Content set 3

For Content set 3, shown in Fig. 4, we found that almost all participants (95 %) produced initial sentences which were in agreement with the ground truth of the situation. This high proportion was not particularly surprising given that only five of the 30 information items in the content set were of low trustworthiness. A notable point here was the occasional inability of individuals to gauge the correct level of an activity or situation. For example, 'looting' was green, but was there a significant amount of looting or one isolated incident of looting? This might also be applied to other cases including whether a situation was very dangerous or not at all dangerous, highlighting some of the limitations of assessing tag clouds. Table 3 shows the words chosen.

The reasons for the words selected for this content set matched several of the previous justifications. For top words such as 'Tottenham', 'police' and 'looter', their size, significant trustworthiness and meaning (i.e., the understanding they added to a situation) all played a key part. The word 'dangerous' (which was used in 2/3rds highly trustworthy content) featured substantially ahead of words such as 'Haringey' (a word only mentioned in highly trustworthy content), report (mostly highly trusted) and officer (mostly highly trusted). One reason quoted was the word's ability to help understand the situation and how dangerous and serious it was. Some participants were also interested in finding out why there was such a clear split between the type of sources mentioning the word, i.e., why were some highly trustworthy yet others not trustworthy.
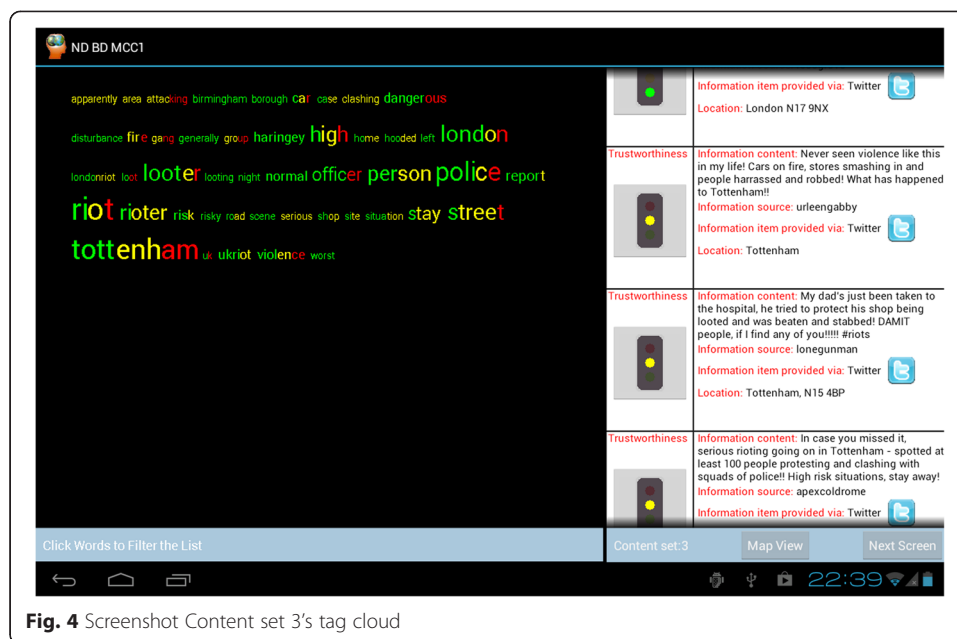
Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 11 of 22

Fig. 4 Screenshot Content set 3's tag cloud

One participant's theory was that maybe one set of sources were saying it was very dangerous and another set saying it was not dangerous – this therefore repeated the point made above.

Participants seemed satisfied with their initial interpretations of the tag cloud, as only 12 % wanted to notably change their summary. More than before, participants' new sentences gravitated towards the highly trustworthy content. Thus, participants had read this information whilst browsing, and were convinced enough to update their sentences. Less significant sentence refinements also resulted in sentences focused on more highly trustworthy content. A potential reason for this might have been the large amount of highly trusted content which led to more substantial impact on their opinions. To compare the full set of sentences (after allowing for updates) with the trustworthy data set content, there was no change in the amount of agreeable sentences.

Table 3 Content set 3: Words selected by participants and their frequencies

Apparently (2)	Normal (6)
Attacking (1)	Officer (4)
Birmingham (1)	Person (1)
Borough (1)	Police (23)
Car (1)	Report (1)
Clashing (1)	Riot (9)
Dangerous (9)	Risk (5)
Disturbance (5)	Scene (1)
Fire (1)	Stay (2)
Haringey (1)	Street (3)
Home (1)	Tottenham (21)
London (4)	Ukriot (2)
Looter (12)	Violence (4)
Looting (1)	Worst (2)

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 12 of 22

Content set 4

For Content set 4, as presented in Fig. 5, we discovered that 85 % of the participants agreed with high and medium trustworthy information content. Errors comparable to those described in previous sections regarding assumptions were apparent in the contingent that did not agree. There were also two cases where 'Birmingham' was wrongly mentioned as the location of an event. Participants were possibly drawn to it because of its use in only medium trustworthy content, but the conclusions they made were not in line with what the medium and high trustworthy content were actually reporting. The words selected are below (Table 4).

As shown above, 'Tottenham' and 'police' were the most selected words and the reasons for their selection matched those in preceding sections (i.e., size, trustworthiness combination, situation understanding and location). Compared to Content set 3, the impact of the varying levels of trustworthiness was clearly visible. For example, there was an increase in the selections of 'attacking', 'street' and 'violence' (to match greater trustworthiness) but a decrease in the choices of 'dangerous', 'disturbance' and 'risk' (to match a decrease in trustworthiness). The changes in frequency of 'risk' and 'street' were especially intriguing as there were only slight changes to the percentage of letters coloured in a specific colour; it was therefore encouraging to see participants recognise and respond to this. This, however, was not always the outcome as with the case of 'Birmingham' (greater trustworthiness in Content set 3 but a slightly greater frequency in set 4). Words that did not change colour much, received roughly the same number of selections, thus alluding to some level of consistency in decisions. 'Tottenham' and 'riot' are examples of this.

There were also fewer participants (only 7 %) that desired to substantially change their initial sentences. Most of the changes made were updating sentences to reflect what individuals read after their browsing of content. Although there was some evidence to support individuals correcting their initial inaccurate statements, unfortunately, some individuals with incorrect statements still maintained their opinion. With

Fig. 5 Screenshot Content set 4's tag cloud

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 13 of 22

Table 4 Content set 4: Words selected by participants and their frequencies

Apparently (3)	Night (2)
Area (1)	Officer (1)
Attacking (6)	Person (3)
Birmingham (2)	Police (20)
Car (1)	Report (1)
Clashing (1)	Riot (11)
Dangerous (2)	Rioter (3)
Disturbance (1)	Risk (1)
Fire (2)	Serious (3)
Gang (3)	Stay (4)
Haringey (1)	Street (7)
London (1)	Tottenham (23)
Loot (1)	Ukriot (1)
Looter (9)	Violence (11)
Looting (1)	Worst (1)

updated sentences considered, 90 % of participants were in agreement with the trustworthy content, demonstrating a reasonable increase given the detailed content they eventually had available.

Content set 5

Unlike the other sets, Content set 5 allowed us to focus and assess how well participants were able to respond to differences in trustworthiness where tense (i.e., either historic – "something was happening", or present – "something is happening") was the key factor. The related screenshot is shown in Fig. 6. From the data, it was apparent that because of this aspect, participants had difficulty in accurately (according to trustworthiness) describing what was occurring. As such, only 51 % of individuals supplied sentences that agreed with high or medium trustworthy content; this was a significant drop compared to other content sets. The issue was not necessarily understanding of the words on screen but the inability to spot that high and medium trustworthy content mentioned that the riots had now ended and clean-up activities were the main on-going activity (Table 5).

Participants largely followed previous thought processes for selecting words. Therefore, 'Manchester', 'riot' and 'shop' were selected for location, size, and reasonable degree of trustworthiness, while 'clean' and 'ended' although small, were key words that conveyed meaning about the scene and had good trustworthiness.

For the follow-up sentence, 32 % of individuals decided to notably change their sentences and in all except two cases, these new sentences were in agreement with trustworthy content. Once participants had the ability to read content, they quickly realised their misconceptions and updated their views accordingly. This finding highlights the potential for the tag cloud interface by itself, to mislead users. For best results, it should be used with care, and actual content (at very least, trustworthy content) referenced for detail of the situation. This increase in understanding could be seen in the revised full set of sentences, where there were 66 % of participants agreeing with trustworthy content. This was a clear increase from the initial 51 %.

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 14 of 22

Fig. 6 Screenshot Content set 5's tag cloud

Content set 6

In Content set 6, as shown in Fig. 7, 78 % of participants reported statements that agreed with trustworthy content. Those who did not supply content in agreement with the ground truth of the situation were split between people that listened to untrustworthy information and people who made assumptions about what was happening in the situation (Table 6).

Regarding the selection of words, typical reasons re-emerged, hence the prevalence of words such as 'Manchester', 'police', 'shop', 'riot' and 'violence'. Comparing this set's selections to those of set 5, differences in trustworthiness were seen to lead to spikes and drops in choices. Words that increased where trustworthiness positively changed included: 'police', 'violence' and 'shop'. Conversely, where trustworthiness in the words 'clean', 'ended' and 'centre' diminished, so did their frequency of selection. What was more interesting however, was the fact that small changes in trustworthiness (i.e., change in the colour of two letters) did have an impact, as seen with the word 'police'; undoubtedly, this also has a link to the size of the word, particularly in this case. 'Shop' also exhibited a similar feature although its change was arguably more profound – moving from 50 % trustworthy to 75 %.

Table 5 Content set 5: Words selected by participants and their frequencies

Boarded (4)	Manchesterriot (2)
Centre (6)	Police (10)
City (2)	Riot (22)
Clean (9)	Shame (2)
Ended (8)	Shop (11)
Justice (3)	Street (1)
Looter (7)	Ukriot (3)
Looting (6)	Violence (2)
Manchester (24)	Youth (1)

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 15 of 22

Fig. 7 Screenshot Content set 6's tag cloud

The results of the second sentence suggest that only 15 % of participants provided completely different sentences, since they were satisfied with their original interpretations. Changes that occurred were simple additions of further detail. Comparing the full set of sentences, only 76 % agreed with high or medium trustworthy content. Although this is a small difference (2 % fewer than the original figure), it could suggest that people had a slightly better grasp of this particular situation using only the tag cloud. This may however, be also linked to a difficulty in finding and believing content where there are only a few items with acceptable degrees of trustworthiness.

Content set 7

For Content set 7, 78 % of participants' sentences agreed with medium or high trustworthy content; the corresponding system screenshot is presented in Fig. 8. Of those sentences that disagreed, there appeared to be a particular misunderstanding of the role that Blackberry played in the scenario. For example, one person commented, "London's Blackberry factory is hit by riots", while another mentioned "Riot concerning Blackberry

Table 6 Content set 6: Words selected by participants and their frequencies

Boarded (1)	Rioting (8)
Centre (1)	Scene (2)
City (2)	Shame (1)
Ended (1)	Shop (18)
Looter (4)	Street (1)
Looting (4)	Uk (1)
Manchester (25)	Ukriot (6)
Police (18)	Video (1)
Report (1)	Violence (13)
Riot (14)	Youth (1)

Nurse *et al. Journal of Trust Management* (2015) 2:10
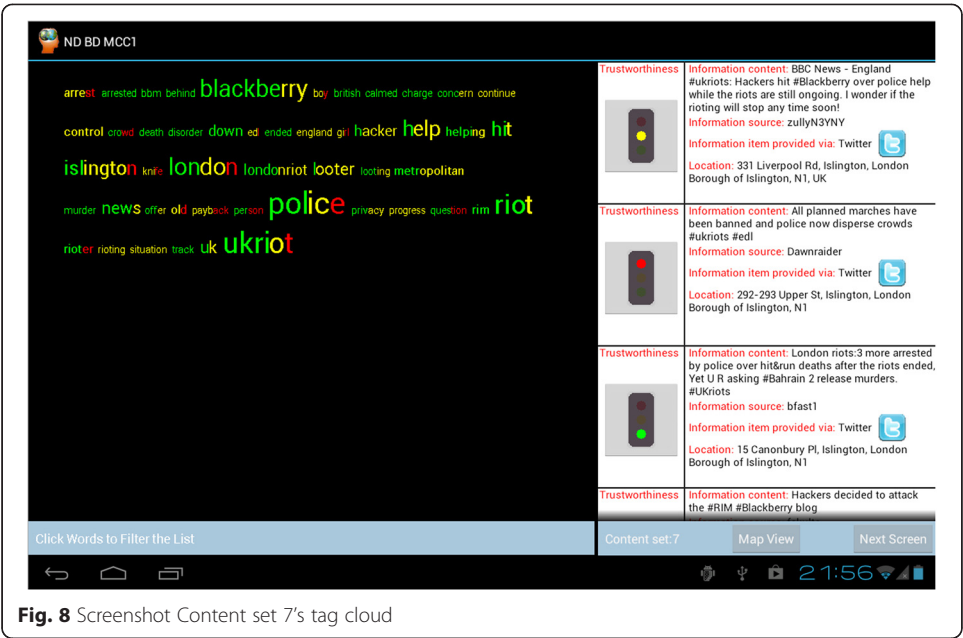
Page 16 of 22

Fig. 8 Screenshot Content set 7's tag cloud

calmed by British police" – actually, 'Blackberry' was assisting the police to track down, catch and arrest rioters and looters. Although the first sentence includes an assumption (i.e., 'factory' was not a word within the cloud), it shows a reasonable thought process as all the other words are in the cloud and are reasonably trustworthy. This is exemplified in the second sentence where the words are all either big or possess a good degree of trustworthiness. Here we see a good example of how the tag cloud might be misread. Below, we present the words selected (Table 7).

Participants appeared to focus heavily on words with either complete trustworthiness (i.e., 'Blackberry' or 'riot') or words that were big, moderately trustworthy and conveyed interesting meaning and insight into the situation (i.e., 'police'). According to participants, 'Blackberry' was also important for them to understand why it was there and how it fit the context of the riot situation. It was interesting that words such as 'murder' and 'death' were not selected that often even though they were only used within highly trustworthy content,

Table 7 Content set 7: Words selected by participants and their frequencies

Arrest (2)	London (4)
Arrested (3)	Looter (2)
Blackberry (32)	Murder (4)
Calmed (3)	News (3)
Charge (1)	Person (1)
Control (2)	Police (22)
Crowd (1)	Progress (1)
Death (2)	Questions (1)
Disorder (1)	Rim (1)
Hacker (4)	Riot (19)
Help (7)	Uk (1)
Hit (1)	Ukriot (7)
Islington (3)	

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 17 of 22

and in the past were viewed as very emotive words that made participants want to find out more. We were not able to investigate this further, and therefore can only hypothesize that when faced with a large number of higher trustworthy and bigger words (or a lesser trustworthiness but still a bigger size), these words may not be that cognitively salient.

Only 15 % of participants provided completely different second sentences. The cloud could therefore be seen as a reasonably good tool at supplying a quick overview that participants were satisfied with even after viewing detailed content. Of the less notable changes made, again, these only sought to add detail to previous perceptions. Considering the change overall, it meant that 90 % of individuals (an increase of 12 % over initial sentences) provided sentences that agreed with the tag cloud and the content specified in the sidebar list.

Content set 8

For Content set 8, presented in Fig. 9, 68 % of participants understood the situation being overviewed. Difficulties arose from misunderstanding how 'Blackberry' was involved. For example, one sentence read, "There is looting/violence in Islington and Blackberry have been hacking people's phones". Actually however, according to trustworthy content, Blackberry was being attacked by hackers. Location also arose as a lesser issue, that is, participants seemed to think that events were primarily occurring in 'Islington', rather than making general comments about 'London'. An understandable reason for this could be that the majority of the word 'Islington' is in green, more so than the word 'London'. Participants might therefore have preferred this word in their sentences (Table 8).

Comparing the words to those from Content set 7, the impact of changes in trustworthiness was clear. Words, 'arrest', 'help', 'Islington' which increased in trustworthiness, all received a higher amount of selections. Whereas, 'Blackberry', 'murder', 'riot' and 'calmed' which had less trustworthiness, elicited a lower selection frequency. Subtle increases in trustworthiness were also reflected to a small degree in word choices, one
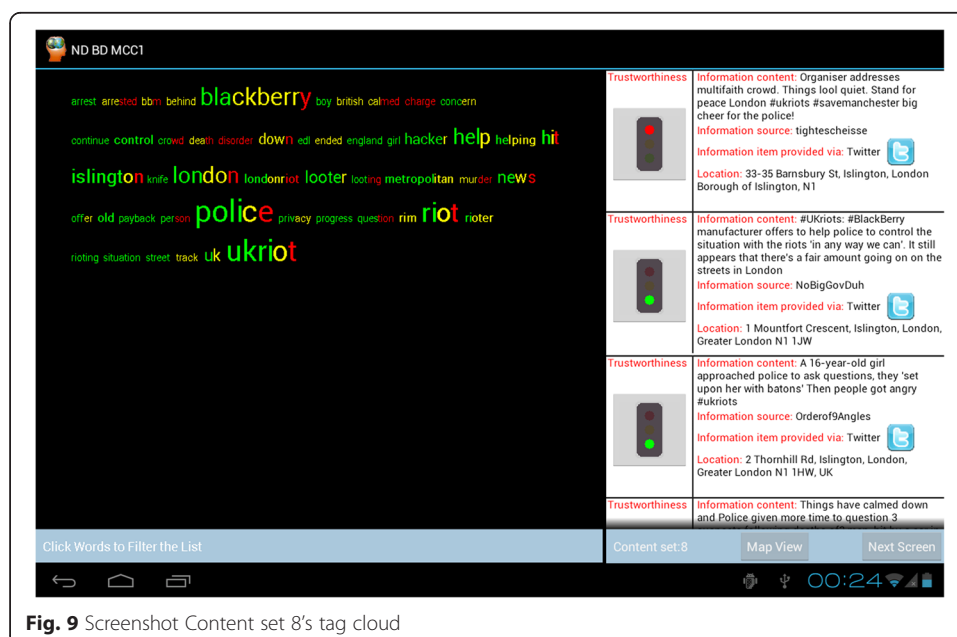

Fig. 9 Screenshot Content set 8's tag cloud

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 18 of 22

Table 8 Content set 8: Words selected by participants and their frequencies

Arrest (5)	Looter (5)
Blackberry (18)	Looting (1)
Concern (1)	Metropolitan (1)
Continue (2)	Murder (1)
Control (4)	News (1)
Crowd (1)	Police (23)
Disorder (1)	Privacy (1)
Hacker (6)	Progress (2)
Help (10)	Riot (2)
Islington (11)	Rioting (2)
Knife (2)	Street (1)
London (3)	Ukriot (8)

example is the word 'police' where one letter was changed from yellow to green which resulted in one increased selection. Of course however, an increase by one in the frequency of selection could be justified by expectable variance or chance, which is why our work focuses more on greater variances.

Only 24 % of participants significantly changed their summaries. Of these, half (5 people) provided sentences that did not fit with high or medium trustworthy content. People that made smaller changes to their content mainly focused on supplementing their original sentences with further detail. Finally, after accommodating for updates in sentences, 71 % of participants supplied overviews that agreed with trustworthy content, showing a slight increase over the 68 % based on only the tag cloud. This percentage was not higher given that individuals had the opportunity to read actual content, but we were not able to ascertain a general hypothesis for this.

Reflecting on the complete experiment

Reflecting on the complete results, there are several noteworthy findings. The first of these pertains to the general utility of the coloured tag cloud and how successful it was in conveying an overview of a scenario. In a majority of content sets, study participants – most of which were new to tag clouds – were able to review the cloud and to feedback coherent summary sentences capturing what was likely to be occurring. This was a significant finding for our research and the use of such clouds in these situations. Moreover, this finding allowed us to answer our main research question positively, which is, that tag clouds with words coloured according to the trustworthiness of related content can facilitate a quick and helpful overview of a scenario.

The analysis of the qualitative data, both from the questionnaire and the interviews, enhanced our belief that tag clouds provide a quick overview of the situation. From the 42 participants who were requested to fill in the questionnaire, 32 responded that the tag cloud was helpful and easy to use. Participants who considered tag cloud as confusing were mainly puzzled by the colouring of the words, sometimes linking the red colour with danger and the green with safety rather than with the degree of trustworthiness of information. This is an observation worth further consideration in the future as it might highlight the potential for misunderstandings with a prospective physical situation awareness tool.

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 19 of 22

During the experiment, we also gathered data to identify and support the importance of features of a word that influenced which words participants focused on and eventually selected (for purposes of gaining detail). These included word size (the bigger the word, the more often it was mentioned in the tweets), trustworthiness colour (words with greater trustworthiness tended to be preferred), meaning (words that appeared to provide more insight into the state of a situation tended to capture people's attention, e.g., 'police' and 'arrest'), and emotion invoked (in some cases, participants found words that described shocking and very serious events, such as murder or killings, worthy of their consideration).

While the size of the word was expected to be clearly understood, it was encouraging to see a notable impact of trustworthiness colour on word selections. This was especially noticeable when comparing content sets where words and their size remained the same, but the trustworthiness of words changed. The meaning and emotion factors were an intriguing finding since they presented a new, but understandable, justification for participants' choices. In retrospect, one can understand accessing a word not necessarily because of its trustworthiness or size, but because it is: (i) insightful and could provide further detail and context for what's happening; or (ii) so serious, that any potential of it occurring would affect a person's decision.

During the interviews, participants also stated that tag clouds helped them focus more on the trustworthy content and avoid being influenced by lower-trustworthiness information. They suggested that by reading all of the information present, there was a possibility that the lower-trustworthiness information would influence their decision even if they tried to focus on the higher-trustworthiness content only. However, the cloud provides a good filtering of the lower-trustworthiness sources because of participants' tendency to focus and tap on higher-trustworthiness words, thus having access only to higher-trustworthiness tweets. Throughout the experiments individuals would select a combination of words based on the size of the words and level of their trustworthiness, but whenever these combinations ended up in lower-trustworthiness tweets they would clear the filter without even reading the remaining tweets and try different combinations.

Our assessment did highlight potential weaknesses and limitations of the tag cloud as a situation overview tool as well. The first of these was apparent in the inability of participants to gauge the amount of an activity (e.g., when people saw the word 'looting', it was hard to determine whether there was a significant amount of looting or an isolated case of looting). Similar problems also could be seen when considering negation (i.e., "rioting is happening in London" versus "there is no rioting in London") and word tense (e.g., "looters are breaking shop windows" versus "looters were present and whilst there, were breaking shop windows"). This inability often resulted in participants misinterpreting the tag cloud and arriving at conclusions that did not agree with trustworthy content.

In the interviews, some participants mentioned that although they believed the tag cloud feature was useful, it had the potential to be misleading. To avoid possible misinterpretations, participants stated that selecting a combination of words and filtering the available tweets would certainly improve the accuracy of their understanding of the situation. In addition, the minority of participants who expressed difficulties in using the clouds stated that their main issue was the uncertainty present with gathering an

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 20 of 22

initial overview of the situation from a tag cloud of coloured words; they expressed a clear preference for reading all the available information.

As a result of the nature of tag clouds and the level of abstraction they provide, regardless of the appearance of the tag cloud and words in it, there exists potential for misinterpretations. This was exemplified perfectly in one content set in particular, where a participant produced a sentence made from reasonably large and mostly highly trustworthy words yet still did not agree with the trustworthy. The problem in this case was not that the important words or context were absent, but rather the way in which they were put together was not correct. This therefore highlights a greater challenge for the application of these tag clouds when used on their own.

Overall, the results show that the coloured tag cloud can be a good and helpful technique at conveying an overview of a situation. There are, however, some limitations of their use which should not be overlooked, since they can lead to severe misunderstandings of the situation. Potentially the best approach – as was undertaken by several participants – is to use the cloud for a quick initial summary of a large data set, but then to reflect on detailed content items when possible, and update one's opinions to suit. At this point, we limit our work to the domain of crisis-management situations, where our messages have a simple three-tiered trustworthiness score. More complex interfaces could be defined to test these findings further, and indeed across other contexts. Lastly, it should be noted that although we aimed for a diverse sample of participants, considering that we relied on volunteers, there may be some sampling bias, for instance, early adopters may have been more likely to sign-up for our experiment. This could mean therefore, that people more likely to explore, adapt and pick-up or understand new things may have led to conclusions regarding our work that may not suit the average user. It would be prudent therefore to rerun this experiment in the future with a more targeted user group with a vested interest in this work and ideally a larger sample.

Conclusions and future work

Technological advances have enabled people to generate, access, and publish vast amounts of content. This overwhelming amount of information cannot be processed in a timely manner, leading to devastating results in cases where situational awareness is of paramount importance. In this paper, we have explored the possibility of using coloured tag clouds to communicate the trustworthiness of information and to facilitate users gaining a quick understating of a situation.

Towards this end, we conducted an experiment where participants were presented with an overwhelming amount of tweets and a tag cloud screen for eight different contexts. The clouds were coloured based on the trustworthiness of the content of the tweets and their sources, and participants were required to provide an overview of the situation in a very short time. It is evident from our results that most of the participants were able to focus on the trustworthy content and provide coherent summations of what was likely to be occurring in each scenario. Thus, tag clouds can facilitate a fast and efficient overview of a situation. In addition, the use of clouds allowed participants to mostly downgrade or even ignore untrustworthy content, which they would otherwise have had to read, and so not to be influenced in their decisions by information of poor quality.

Future work could focus on exploring different ways of better communicating the trustworthiness of information, since some participants suggested that the colouring

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 21 of 22

scheme for trustworthiness may be confusing on occasion. Future research could shed light on how to disambiguate these situations by also focusing on how to merge synonymous words and how to present words in a manner that would better explain the situation (i.e., adjacent words in a tag cloud would indicate that these appear in the same sentence in some tweets). We would also endeavour to extend our work by conducting another experiment with experts in managing crisis situations, such as police and fire officers (similar to our work in [25]), to explore whether clouds can assist them in their duties by providing quick and reliable situational awareness. This would be a targeted user group and would also assist us in avoiding issues such as sampling bias and small sample sizes.

Another topic that could be explored in the future is whether more detailed classes of trust values (i.e., more than three) could be mapped to a coloured tag cloud and how might that be achieved. Expanding the mapping would allow stratified schemes with more classes to be considered (e.g., 5 star ratings) and therefore, allow a wider application of the clouds (e.g., possibly to review sites). We will need to be careful however, because more levels (and colours) may confuse users and risk being more of a hindrance than a benefit. In the future, we will consider other techniques but will be guided by usability research and exploratory pilot studies as to the best approach that could deal with multiple trust classes if they exist.

Author details
[1]Cyber Security Centre, Department of Computer Science, University of Oxford, Oxford, UK. [2]Department of Psychology, University of York, York, UK.

References
1. Gomez-Rodriguez M, Gummadi KP, and Scholkopf B (2014) Quantifying Information Overload in Social Media and its Impact on Social Contagions. In: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media. AAAI. pp. 170–179
2. ESPN.com. (2013) Boston bombings led 2013 tweets. http://espn.go.com/boston/nba/story/_/id/10124297/boston-marathon-bombings-most-tweeted-sports-event-2013. Accessed 13 April 2015
3. Time Inc. (2014) Fear, Misinformation, and Social Media Complicate Ebola Fight. http://time.com/3479254/ebola-social-media. Accessed 13 April 2015
4. Heimerl F, Lohmann S, Lange S, and Ertl T (2014) Word cloud explorer: Text analytics based on word clouds. In: Proceedings of the 47th Annual Hawaii International Conference on System Sciences (HICSS). IEEE. pp. 1833–1842
5. Nurse JRC, Creese S, Goldsmith M, and Lamberts K (2012) Using Information Trustworthiness Advice in Decision-Making. In: Proceedings of the International Workshop on Socio-Technical Aspects in Security and Trust (STAST 2012) at the 25th IEEE Computer Security Foundations Symposium (CSF). IEEE. (pp. 35–42)
6. Nurse JRC, Creese S, Goldsmith M, and Rahman SS (2013). Supporting Human Decision-Making Online Using Information-Trustworthiness Metrics. In Human Aspects of Information Security, Privacy, and Trust. Springer Berlin Heidelberg. (pp. 316–325)
7. Nurse JRC, Agrafiotis I, Creese S, Goldsmith M, and Lamberts K (2013) Building Confidence in Information - Trustworthiness Metrics for Decision Support. In: Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-13). IEEE
8. Nurse JRC, Agrafiotis I, Creese S, Goldsmith M, and Lamberts K (2014) Two sides of the coin: measuring and communicating the trustworthiness of online information. Journal of Trust Management. Springer. 2014, 1:5.

Nurse *et al. Journal of Trust Management* (2015) 2:10

Page 22 of 22

9. Keeton K, Mehra P, Wilkes J (2010) Do you know your IQ?: A research agenda for information quality in systems. SIGMETRICS Performance Evaluation Review 37(3):26–31

10. Kelton K, Fleischmann KR, Wallace WA (2008) Trust in digital information. J Am Soc Inf Sci Technol 59(3):363–374

11. Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. J Manag Inf Syst 12(4):5–33

12. Rieh S and Belkin N (1998) Understanding judgment of information quality and cognitive authority in the WWW. In: Proceedings of the 61st ASIS Annual Meeting, pp. 279–289

13. Visschers VH, Meertens RM, Passchier WW, De Vries NN (2009) Probability information in risk communication: a review of the research literature. Risk Anal 29(2):267–287

14. Lipkus I, Hollands J (1999) The visual communication of risk. JNCI Monographs 25:149–163

15. Lipkus I (2007) Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. Med Decis Mak 27(5):696–713

16. Chevalier F, Huot S, and Fekete J-D (2010) Wikipediaviz: Conveying article quality for casual Wikipedia readers. In: Proceedings of the Pacific Visualization Symposium (PacificVis). IEEE. pp. 49–56

17. Idris NH, Jackson MJ, and Abrahart RJ (2011) Colour coded traffic light labeling: A visual quality indicator to communicate credibility in map mash-up applications. In: Proceedings of International Conference on Humanities Social Sciences, Science & Technology

18. Adler BT, Chatterjee K, De Alfaro L, Faella M, Pye I, and Raman V (2008) Assigning trust to Wikipedia content. In: Proceedings of the 4th International Symposium on Wikis. ACM

19. Volk F, Hauke S, Dieth D, and Muhlhauser M (2014) Communicating and visualising multicriterial trustworthiness under uncertainty. In: Proceedings of the 12th Annual International Conference on Privacy, Security and Trust (PST). IEEE. pp. 391–397

20. Seifert C, Kump B, Kienreich W, Granitzer G, and Granitzer M (2008) On the beauty and usability of tag clouds. In: Proceedings of the 12th International Conference on Information Visualisation (IV). IEEE. pp. 17–25

21. Kuo BY, Hentrich T, Good BM, and Wilkinson MD (2007) Tag clouds for summarizing web search results. In: Proceedings of the 16th International Conference on the World Wide Web. ACM. pp. 1203–1204

22. Hearst MA, and Rosner D (2008) Tag clouds: Data analysis tool or social signaller? In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS). IEEE. pp. 160–170

23. Halvey MJ, and Keane MT (2007) An assessment of tag presentation techniques. In: Proceedings of the 16th International Conference on World Wide Web. ACM. pp. 1313–1314

24. The Guardian (2011) UK riots 2011 http://www.theguardian.com/uk/london-riots. Accessed 13 April 2015

25. Nurse JRC, Creese S, Goldsmith M, Craddock R and Jones G (2012) An Initial Usability Evaluation of the Secure Situation Awareness System. In: Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM). ISCRAM