Combating Misinformation on the Social Media

Xiuzhen Jenny Zhang, (xiuzhen.zhang@rmit.edu.au)

RMIT University

RMIT UNIVERSITY

What is misinformation?

According to the Oxford English Dictionary, misinformation is “false or inaccurate information, especially that which is deliberately intended to deceive”.

- Disinformation
- False rumour
- Spam
- Fake news

Social media is a double-edged sword …

Misinformation on Twitter

Opinion spam

- Opinion spam can range from annoying self-promotion of an unrelated website or blog to deliberate deceptive reviews.
- Deceptive opinion spam – fictitious opinions that are deliberately written to sound authentic to deceive readers – is not easily identifiable by humans.

Which is a spam review?

Review 1:

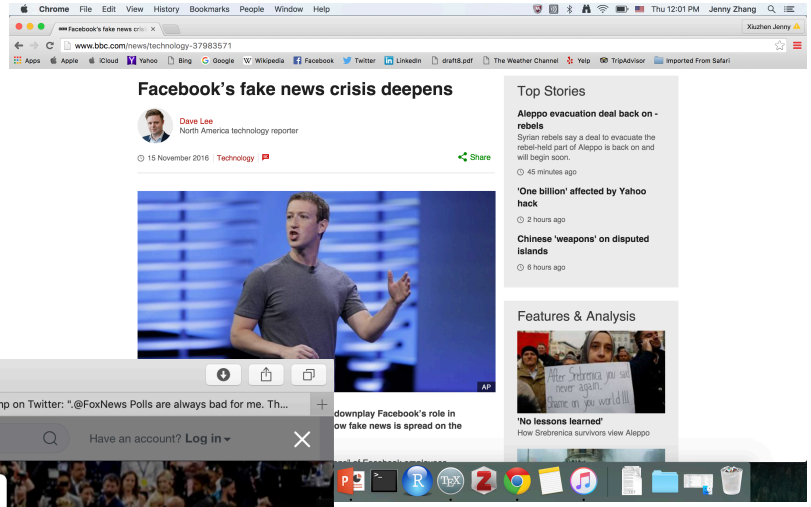
I have stayed at many hotels traveling for both business and pleasure and I can honestly stay that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.

Review 2:

My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop. I couldn't ask for more!! We will definatly be back to Chicago and we will for sure be back to the James Chicago.

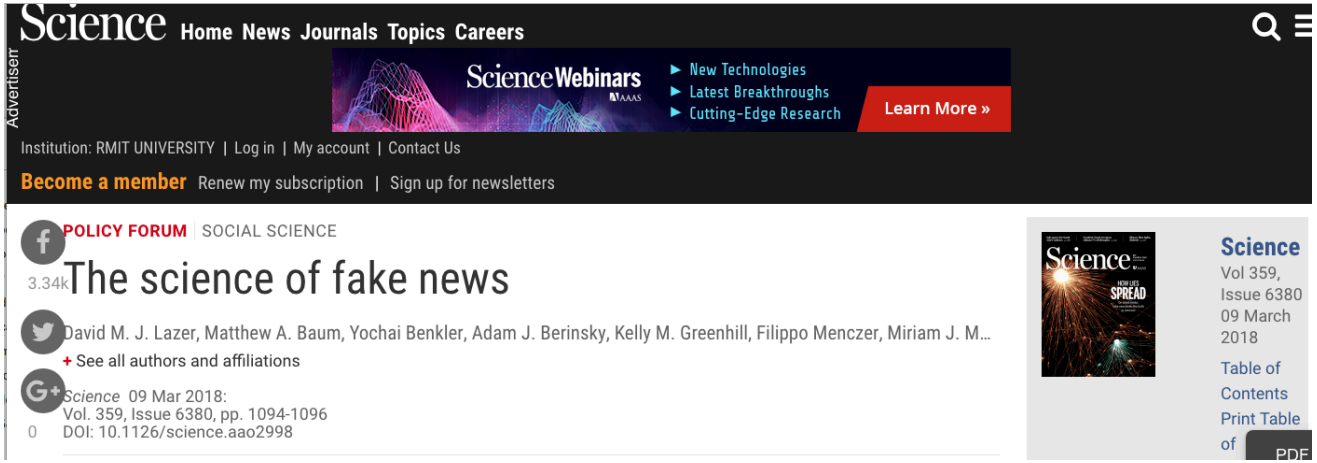
Note: Example taken from (Ott et al, ACL 2011).

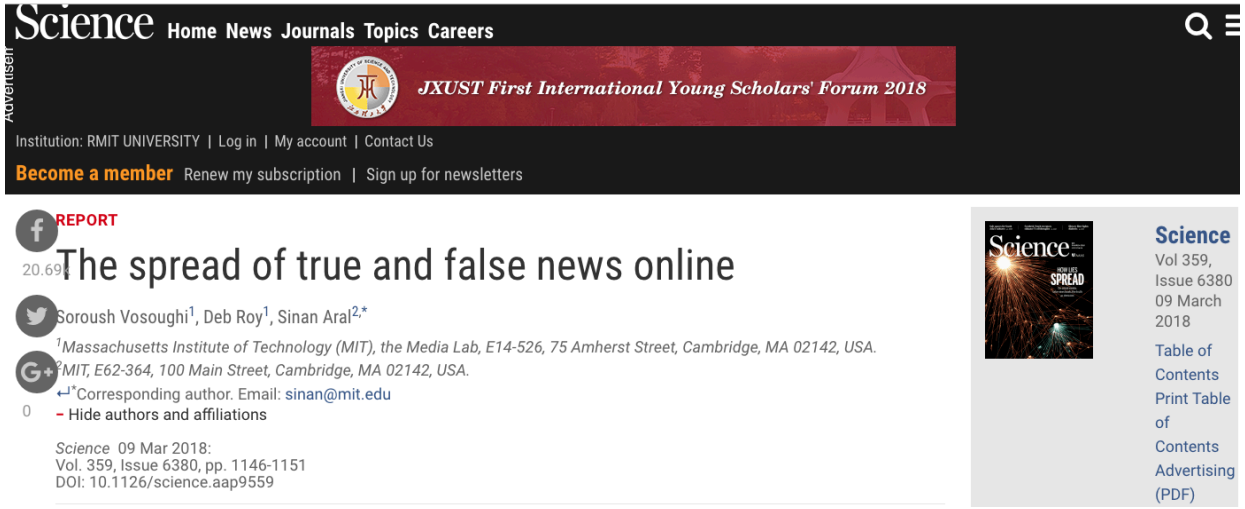
Fake news


Facebook's fake news crisis deepens
Dave Lee
North America technology reporter
Top Stories
Aleppo evacuation deal back on - rebels
'One billion' affected by Yahoo hack
Chinese 'weapons' on disputed islands
Features & Analysis
'No lessons learned'
How Srebrenica survivors view Aleppo


Donald J. Trump
@realDonaldTrump
Follow
.@FoxNews Polls are always bad for me. They were against Crooked Hillary also. Something weird going on at Fox. Our polls show us leading in all 17 Swing States. For the record, I didn't spend 30 hours with @abcnews, but rather a tiny fraction of that. More Fake News @BretBaier
4:49 PM - 17 Jun 2019
16,007 Retweets 64,550 Likes
20K 16K 65K
© 2019 Twitter About Help Center Terms Privacy policy Cookies Ads info
Donald J. Trump
@realDonaldTrump
45th President of the United States of America
Washington, DC
Instagram.com/realDonaldTrump
Joined March 2009

Misinformation spread quicker, farther, and deeper on the social media





But, people are trusting online information

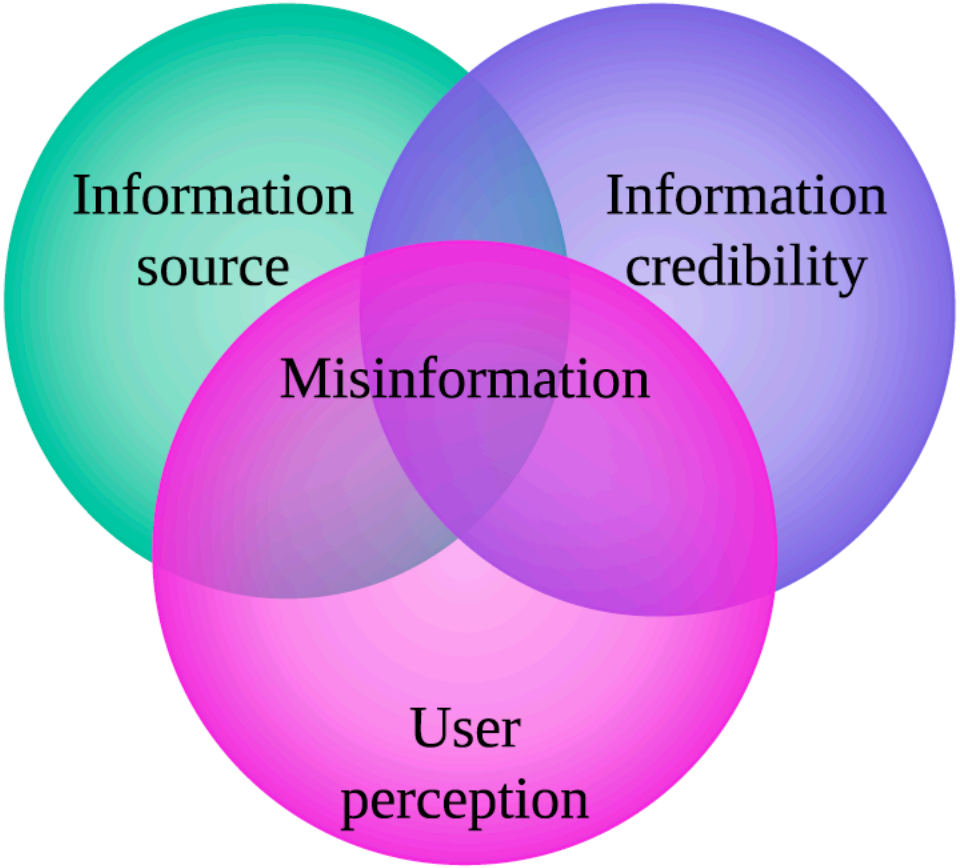
Students Have 'Dismaying' Inability To Tell Fake News From Real, Study Finds

November 23, 2016 · 12:44 PM ET

CAMILA DOMONOSKE

So, research is urgently needed to combat misinformation spreading on the social media.

Misinformation: a holistic view

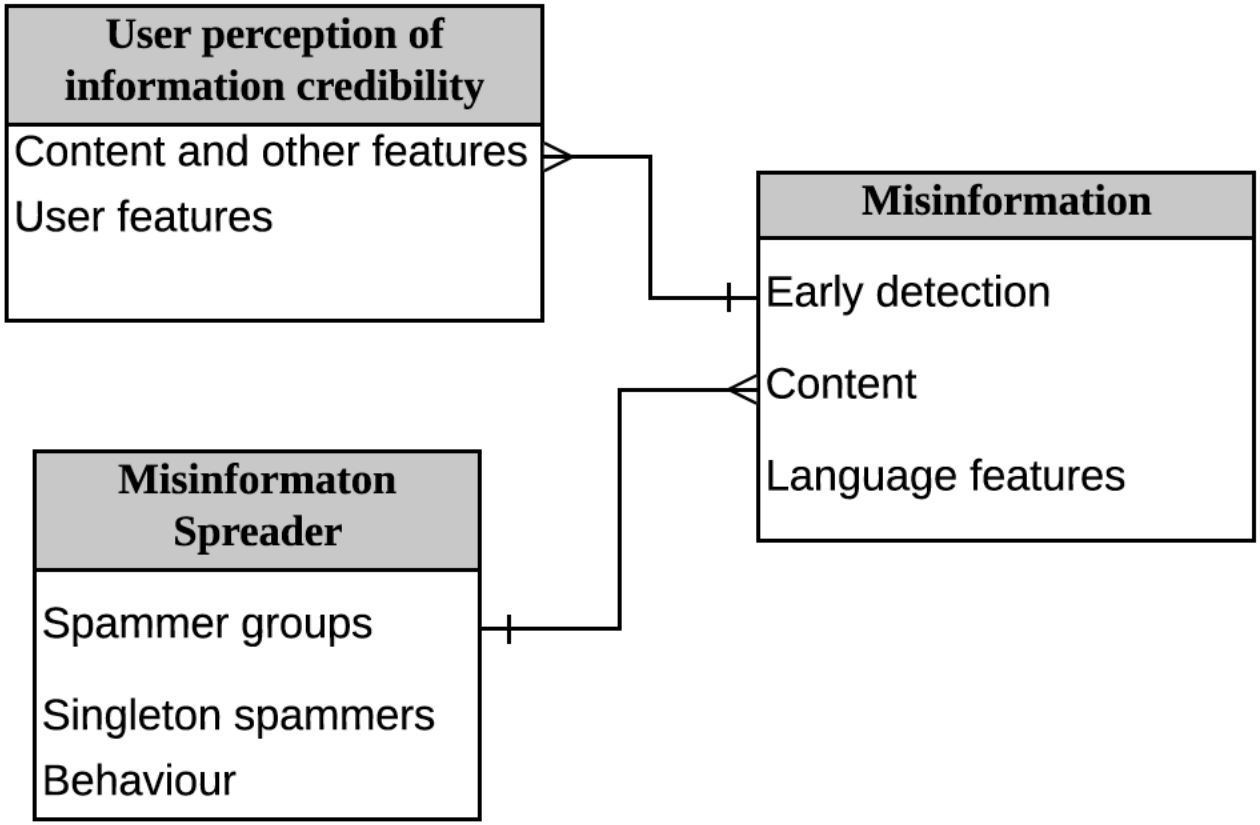


What is credibility?

- Credibility – *the quality of being believed or accepted as true, real, or honest* (The Merriam Webster dictionary).

On the five-level rating scale, seven human judgements have six 'very credible' judgements and one 'seem credible' judgement.

Two perspectives of misinformation

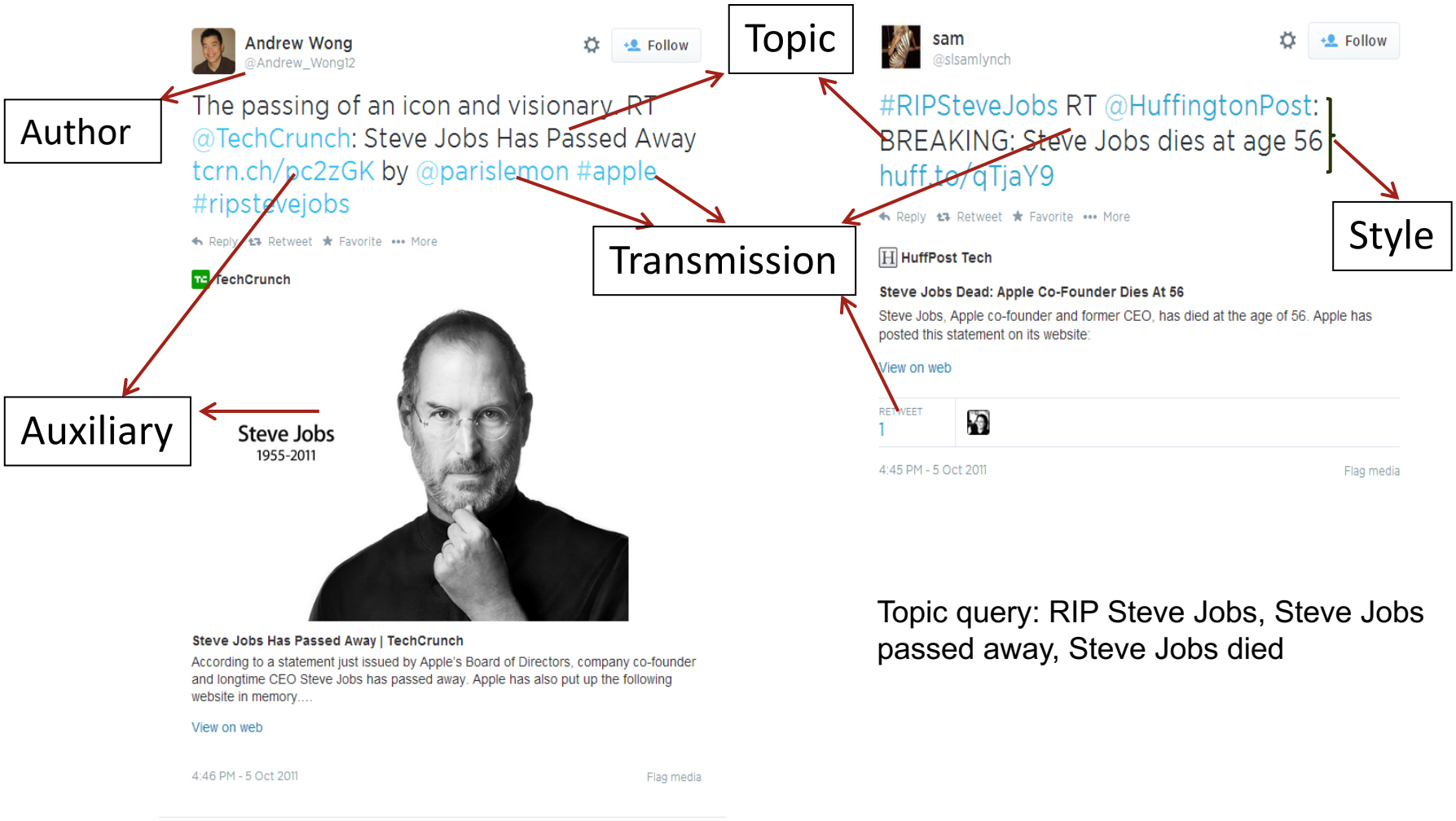


I. User perception of information credibility

User Perception of Information Credibility studies using a crowdsource platform

- Tweet collection: tweets of four years on 60 news topics Science, Politics, and Natural Disaster.
- Tweets are mixed true and fake news posts.
- Crowdsource platform figure-eight.com was used to collect human judgements of credibility level of tweets.
- The demographic and location information of users was collected.
- ANOVA and Chi-square analysis.

A user study on features for users to judge credibility of news on Twitter



Topic query: RIP Steve Jobs, Steve Jobs passed away, Steve Jobs died

Findings – tweet contents

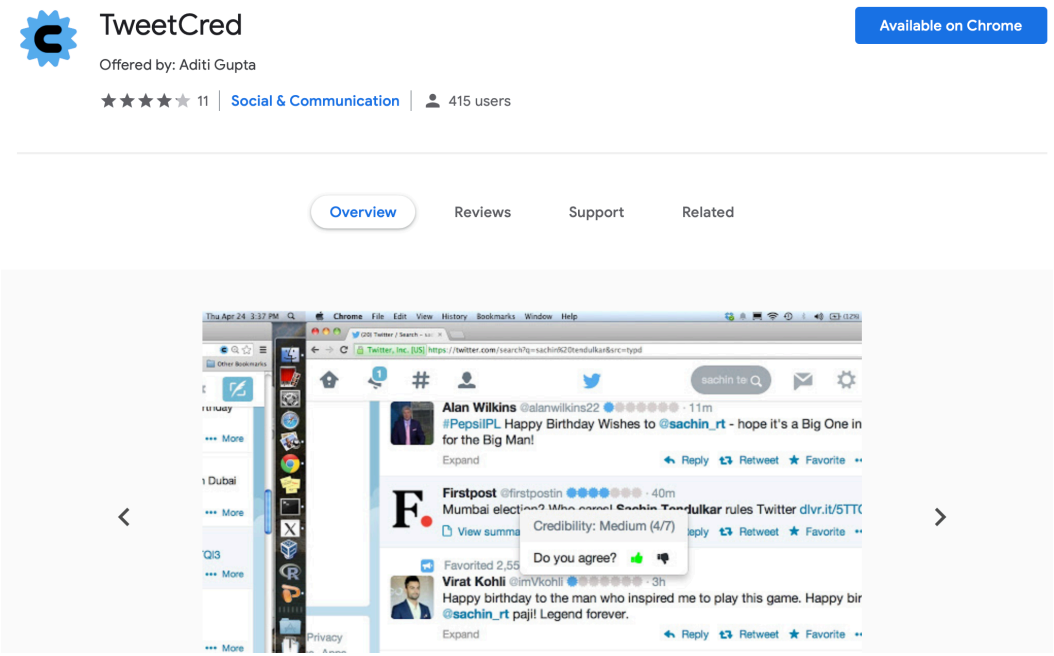
- A large portion of readers judge tweet credibility based on topic and style features.
- Auxiliary, Transmission and Author features are mostly combined with other features by users (readers) to judge tweet credibility.

Findings – users

- The demographic information of users is strongly correlated with their perception of information credibility on Twitter.
- Location of users affects their perception of information credibility.

Who is more trusting? Humans or machines?

- TweetCred – a machine learning tool for prediction of information credibility for tweets.



Findings – humans vs. machines.

Humans are more trusting of news tweets than machine prediction.

Table 4.4: The agreement matrix between reader's credibility perception and automated credibility prediction

		TweetCred			
		Very credible	Somewhat credible	Not credible	Total
Readers	Very Credible	256	654	67	977
	Somewhat credible	51	230	50	331
	Not credible	1	4	3	8
	Total	308	888	120	1316

Shariff, Shaza Mohd, Xiuzhen Zhang, and Mark Sanderson, 2017. On the Credibility Perception of News on Twitter: Readers, Topics and Features. *Computers in Human Behavior*, Vol. 75, pp. 785-796.

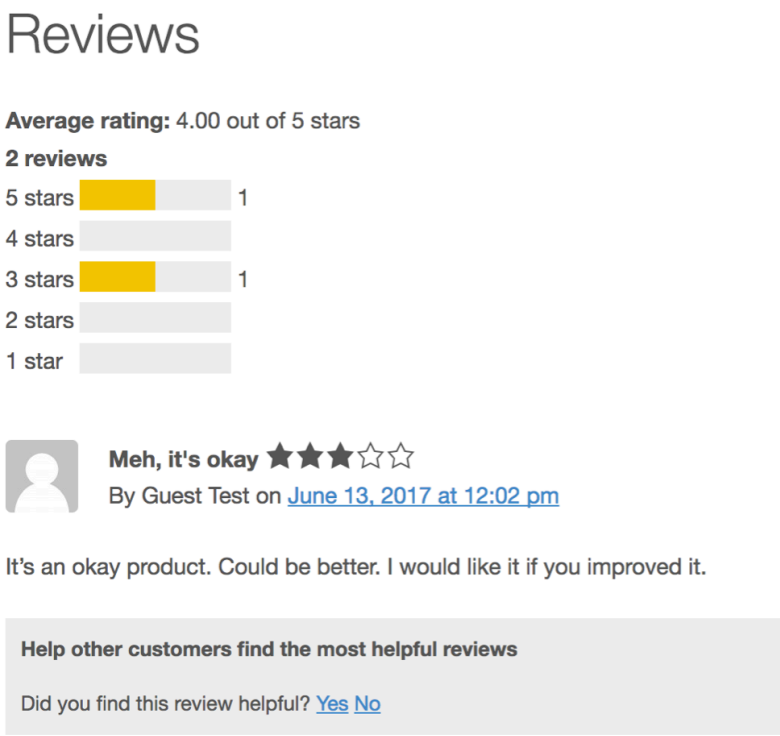
II. Prediction of misinformation

Detection of misinformation

- Detection of misinformation on the social media using media contents only is challenging:
  - Misinformation spreaders make contents look genuine to deceive readers.
- Use a range of information to detect misinformation:
  - The credibility of information source.
  - The behavior trail of information source
  - The social network of information sources

Detection of opinion spam based on anomalous rating deviation

- How to detect spam reviews when text information is not available?
- Spammer manipulate average star ratings for products.
- How to find the true majority opinion in the presence of spam ratings?



Detection of opinion spam based on anomalous rating deviation …

- Our method uses binomial regression to identify reviewers having an anomalous proportion of ratings that deviate from the majority opinion.
- To compute the true mean rating as the majority opinion, we apply an iterative process whereby the contribution from each reviewer to the average rating for each product is successively reduced based on their proportion of non-majority reviews.

Findings

- The proportion of reviews disagreeing with the mean is a good indicator of spammer behaviour.
- A main advantage of our approach is its simplicity and the consequent minimal computational requirements.

Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2015). Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications*, *42*(22), 8650-8657.

Identifying singleton review spammers

- Singleton spammers write once-off spam review. There are not any behavior trails.
  - There are many sock puppet accounts on review sites.
- But to make their spam campaign effective, they form spammer groups and issue coordinated attacks.

Identifying singleton spammers …

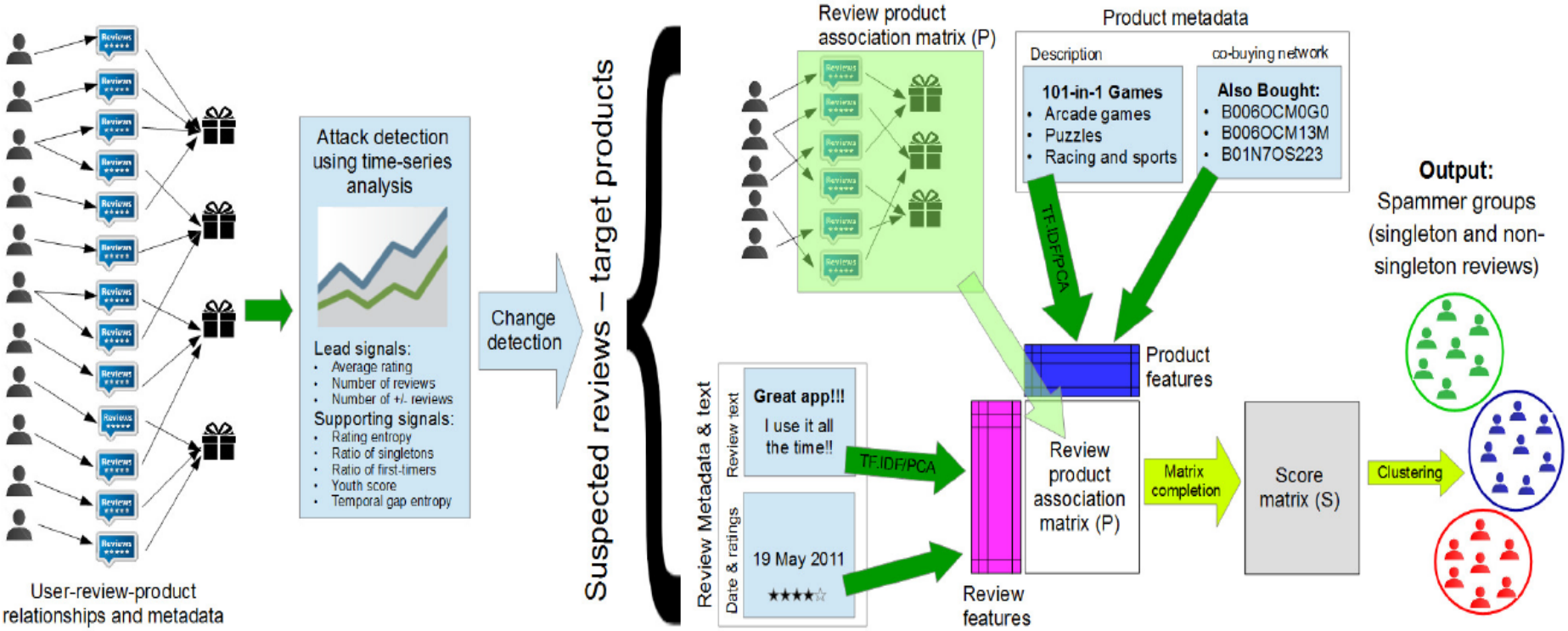


Fig. 1. Schematics of the proposed approach (SSGD)

The review-product graph defined by attacks

- The review-product bipartite graph for suspicious reviews and target products is a review-product association matrix.
- Challenge: The review-product graph is sparse. How to uncover the hidden reviewer-product associations?

Approach: IMC (Inductive Matrix Completion)

- The IMC algorithm was used to predict gene-disease associations by combining multiple types of evidence (features) for diseases and genes to learn latent factors that explain the observed gene-disease associations.
- IMC assumes that the associations matrix is generated by applying feature vectors associated with its rows and columns to a low-rank matrix *Z*, and solve for *Z*.

Natarajan, N., Dhillon, I.S.: Inductive matrix completion for predicting gene-disease associations. *Bioinformatics* 30(12), i60-i68 (2014)

Findings

Detecting singleton spammers via uncovering their hidden collusiveness and groups is an effective strategy for detecting singleton spammers.

- Spammer groups detected mostly consist of singleton reviewers give all high (4-5 star) or all low (1-2 star) ratings and write nearly identical reviews within a short period of time.
- A group typically consists of 20-90 reviewers (many are singletons) targeting 4-9 products.
- The timestamps and rating distribution of most groups are concentrated for spam attacks.

Kumar, D., Shaalan, Y., Zhang, X., & Chan, J. (2018). Identifying singleton spammers via spammer group detection. In *Proc. PAKDD 2018*

III. Future work

Looking ahead: Deep learning for misinformation detection

- It is desirable to design deep learning models that can extract deeper, hidden information from various signals, e.g., information source and propagation paths, contents.
  - Early detection
  - User embedding and content embedding
  - Lack of annotated resources
  - How to incorporate heuristics by human experts?

Looking ahead: mitigation of misinformation

- Humans are trusting online information and cause the spread of misinformation.
- Rather than post-hoc detection, proactive mitigation is most desirable.
  - Optimization for true information spread over the social network to mitigate misinformation.
  - Personalised recommendation of true information to combat misinformation.

Further information

- Shaalan, Y., Zhang, X., Chan. Detecting Singleton Spams via Deep Mining for Anomalous Temporal Aspect-Sentiment Patterns. *Proc. ICDM 2019*. In submission.
- Aladhadh, S., Zhang, X., & Sanderson, M. (2019). Location impact on source and linguistic features for information credibility of social media. *Online Information Review*, *43*(1), 89-112.
- Pourhabibi, T., Boo, Y. L., Ong, K. L., Kam, B., & Zhang, X. (2018). Behavioral Analysis of Users for Spammer Detection in a Multiplex Social Network. In *Proc. AusDM 2018*.
- Kumar, D., Shaalan, Y., Zhang, X., & Chan, J. (2018). Identifying singleton spammers via spammer group detection. In *Proc. PAKDD 2018*.
- Shariff, S. M., Zhang, X., & Sanderson, M. (2017). On the credibility perception of news on Twitter: Readers, topics and features. *Computers in Human Behavior*, *75*, 785-796.
- Aladhadh, S., Zhang, X., & Sanderson, M. (2017). Beyond the culture effect on credibility perception on microblogs. In *Proc. SocInfo 2017*.
- Shariff, S. M., Sanderson, M., & Zhang, X. (2016). Correlation analysis of reader's demographics and tweet credibility perception. In *Proc. ECIR 2016*.
- Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2015). Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications*, *42*(22), 8650-8657.
- Aladhadh, S., Zhang, X., & Sanderson, M. (2014). Tweet author location impacts on tweet credibility. In *Proc. ADCS 2014.*
- Shariff, S. M., Zhang, X., & Sanderson, M. (2014). User perception of information credibility of news on Twitter. In *Proc. ECIR 2014.*

Thank you

- For this opportunity to share our research
- Acknowledgements:

Australian Government
Australian Research Council

CSIRO DATA 61

For more information …

Contact: xiuzhen.zhang@rmit.edu.au

http://www.xiuzhenzhang.org/