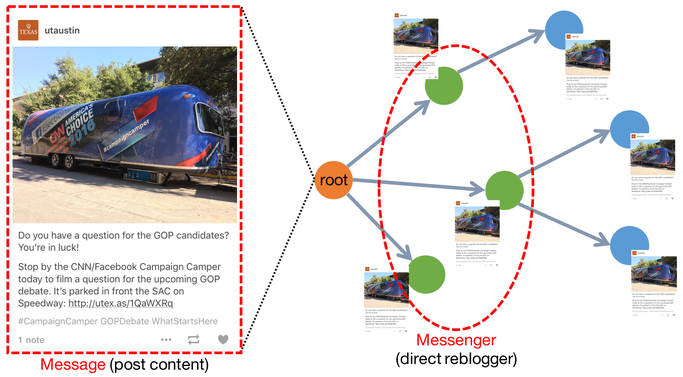
Message or Messenger: An Empirical Analysis on Tumblr Reblogging Behavior

Donghyuk Shin[1], Shu He[1], Gene Moo Lee[2] and Andrew B. Whinston[1]
[1]University of Texas at Austin [2]University of Texas at Arlington

Introduction

- Effectively delivering messages on social media — an important issue for online advertising/marketing.
- Two main aspects to consider:
  - Message: articulating the content in the right format.
  - Messenger: selecting the right initial information injection points to maximize information diffusion.



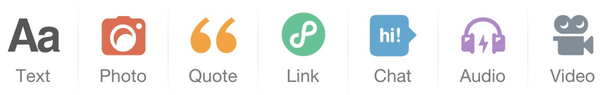
GOAL: *Empirical analysis of sharing behavior on online social media.*

- Focus on messages created by companies.
- Utilize both textual and visual semantic content features obtained using state-of-the-art machine learning methods.

Tumblr

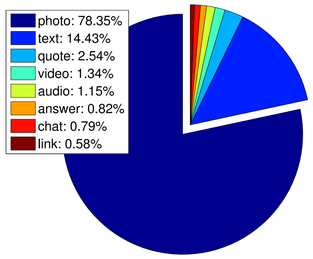
MICROBLOGGING SERVICE:

- Supports various types of posts (no limit on length of posts):



SOCIAL NETWORK SERVICE:

- Follow other blogs — follower graph
- Like or reblog other posts



Hypotheses Development

- **HYPOTHESIS 1 (H1)**: *Company posts with photos or videos will receive more reblogs.*
- **HYPOTHESIS 2 (H2)**: *Posts with less complex photos will receive more reblogs.*
- **HYPOTHESIS 3 (H3)**: *Posts reblogged by users who have more followers will receive more reblogs.*
- **HYPOTHESIS 4 (H4)**: *Posts reblogged by users whose previous posts have high similarity with the focal posts will receive more reblogs.*

Feature Construction

TUMBLR DATASET:

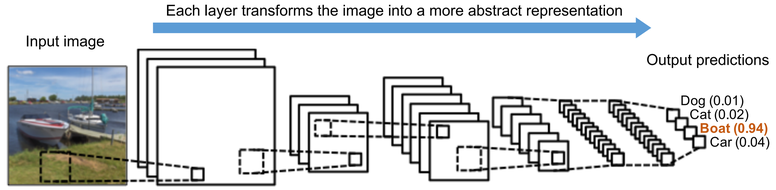
- 3,773 posts created by 102 *official company blogs* over a month.
- 1.8M posts created by 168K direct rebloggers of the company posts.
- More than 1M images collected from posts.

TOPIC MODELING FOR TEXT AND TAGS:

- Latent Dirichlet Allocation (LDA) topic modeling:
  - Document consists of a few latent topics and words in the document is the realization of the underlying topics. [Blei et.al, JMLR 2003]
  - Successfully adopted for many tasks in the management literature.
- Two text corpora: Tags and Text
- Learned 50-topics — we found that the topics are good representations: `http://diamond.mccombs.utexas.edu/lda-tumblr-tag-50-topics.txt`

DEEP LEARNING FEATURES FOR IMAGES:

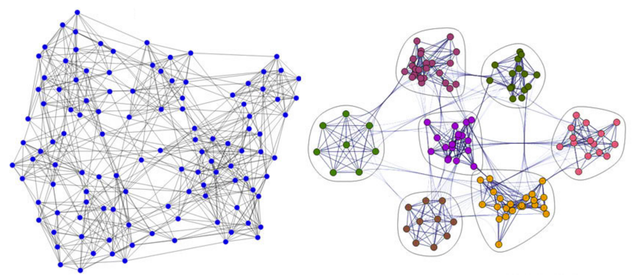
- Majority of posts are *photos* ($\approx$ 80%).
- Traditionally, processing raw visual data required significant engineering efforts and domain knowledge.
- Recent breakthrough of deep learning methods have emerged as powerful models to learn useful image representations from data.
  - Dramatically improved state-of-the-art performance on recognition tasks.



- Employed a deep Convolutional Neural Network (CNN) to extract useful features from the 1M images. [Krizhevsky et.al. NIPS 2012]
- Output is a confidence score vector $p \in [0,1]^{1000}$ corresponding to 1,000 object categories.

COMMUNITY INFORMATION:

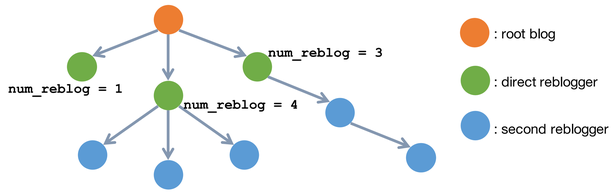
- Utilize community structure of the follower graph.
- Adopted modularity optimization to identify tightly connected communities.
  - *Modularity*: difference between the expected and actual number of intra-cluster edges. [Newman and Girvan, Phys. Rev. E. 2004]
  - From the follower graph with 76.86M nodes and 2.27B edges — found 2,895 communities.



Community detection using modularity optimization

Variables

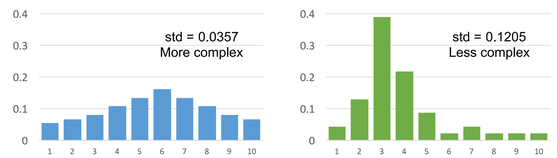
- **TARGET VARIABLE**:
  `num_reblog`: total number of reblogs received through a *direct reblogger*.



- **INDEPENDENT VARIABLES**:

Variable name	Description
`num_reblog`	Number of reblogs received by a direct reblogger
`{feat}_num`	Number of `{feat}` in company post*
`{feat}_sd`	Standard deviation of `{feat}` for company post*
`{feat}_sim`	Cosine similarity of `{feat}` between company post and blog*
`in/out_degree`	Log-scale in/out degree of user blog
`gif, videos`	Company post contains gif/video
`community`	⟨company, user⟩-pair belongs to the same community
`if_follower`	User follows company blog

*`feat` = `{image(photos),text(words),tags}`



Empirical Results

Variable	Negative Binomial	Neg. Binomial - interaction	Negative Binomial	Logit
`video`	-0.13(0.10)	0.73(0.63)		
`video_int`		-0.09(0.07)		
`gif`	0.07(0.05)	1.56(0.38)**		
`gif_int`		-0.16(0.04)**		
`photo_num`	0.67(0.009)**	0.38(0.07)**		
`photo_int`		-0.03(0.008)**		
`tags_num`	-0.03(0.007)**	-0.04(0.007)**		
`words_num`	0.002(0.002)	0.001(0.001)		
`image_sd`	13.99(2.18)**	13.02(2.18)**		
`text_sd`	-0.38(0.39)	-0.29(0.39)		
`tags_sd`	0.63(0.37)	0.80(0.37)*		
`image_sim`	-0.037(0.075)	-0.025(0.075)	-0.055(0.063)	0.18(0.084)*
`text_sim`	0.001(0.08)	-0.027(0.08)	0.48(0.039)**	0.43(0.054)**
`tags_sim`	0.49(0.075)**	0.80(0.369)**	0.40(0.03)**	0.17(0.044)**
`in_degree`	0.19(0.017)**	2.18(0.26)**	0.69(0.003)**	0.72(0.005)**
`out_degree`	0.01(0.014)	0.01(0.014)	-0.086(0.03)**	-0.093(0.004)**
`community`			0.17(0.011)**	0.12(0.014)**
`if_follower`			0.12(0.02)**	0.061(0.023)**
Constant	-2.36(0.17)**	-2.98(0.21)**	-4.87(0.032)**	-3.90(0.106)**
Fixed Effect	Company blog		Company post	
Observations	3,437		267,807	

- **H1**: More photos (`photo_num`) yield more reblogs.
- **H2**: Conspicuous images (`image_sd`) have positive impact.
- **H3**: Rebloggers with more followers (`in_degree`), but less followees (`out_degree`) get more reblogs.
- **H4**: More reblogs when reblogged by users with similar content (`*_sim`).

Conclusion

- Our approach for semantic content analysis, particularly for visual content, bridges advanced machine learning techniques for effective marketing and social media strategies.

Email: dshin@cs.utexas.edu, shuhe@utexas.edu, gene.lee@uta.edu, abw@uts.cc.utexas.edu