



# **CS834 - Introduction to Information Retrieval**

## **Presentation #5**

Hussam Aldeen Hallak

# The Two Papers:

- Finding high-quality content in social media

WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining  
Pages 183-194

- How opinions are received by online communities: a case study on amazon.com helpfulness votes

WWW '09 Proceedings of the 18th international conference on World wide web  
Pages 141-150

# The First Paper:

- Finding high-quality content in social media

WSDM '08 Proceedings of the 2008 International Conference  
on Web Search and Data Mining  
Pages 183-194

<http://www.mathcs.emory.edu/~eugene/papers/wsdm2008quality.pdf>

# Web Transformation

- **1990s**

Traditional published material  
Content created by publishers  
Web users are consumers

- **2000s**

User generated content became popular

Examples:

Blogs

Forums

Photo and video sharing communities: Youtube

Social networks: Facebook

**Question / Answer communities: Yahoo! Answers**

# Yahoo! Answers

- Users ask/answer questions on any topic
- Users regulate the system through voting:
  1. Mark a question as “Interesting”
  2. Mark an answer as “Best answer”
  3. Vote on Answers “Thumbs up/down”
  4. Report abusive content/behavior
- Users have threefold role:
  1. Asker
  2. Answerer
  3. Evaluator
- Users are forming a social network characterized by heterogeneous interactions

# Question Lifecycle

- Open: The question is asked by the user
- Other users answer the question
- Closed: The question gets closed because:
  1. Time limit has been reached (4 days by default)
  2. Asker closed the question
- Resolved: The question gets marked as resolved because:
  1. Asker selected the best answer
  2. Other users voted on the best answer



# Goal

- Identify high quality content in community-driven question/answering sites
- Examine quality range in user-generated content, which is much wider than traditional content
- Utilizing additional sources in the content
  1. Link analysis
  2. User-to-document relations
  3. User-to-User relations

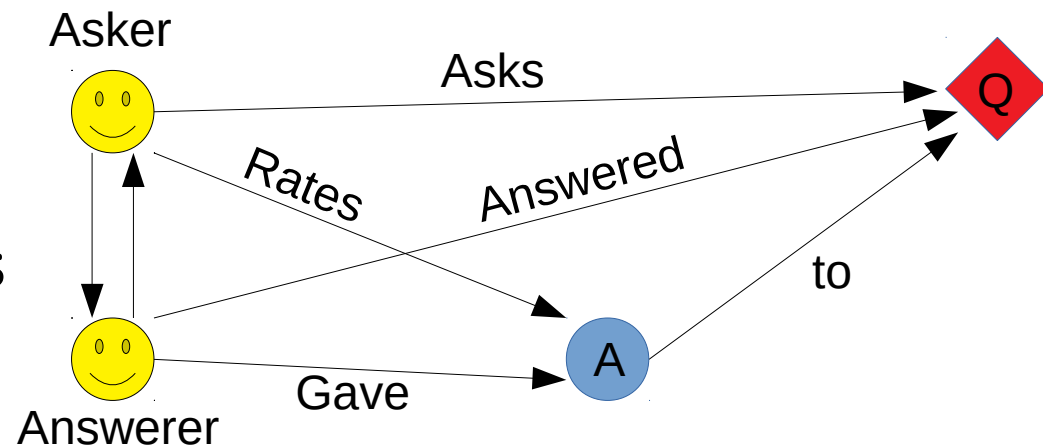
# Content Quality Analysis in Social Media

- Intrinsic content quality: Mostly text-related

1. Punctuation and typos
2. Syntactic and semantic complexity
3. Grammaticality

- User relationships:

1. Nodes: Entities  
users, questions, answers, ...etc
2. Edges: Actions & relationships  
answers, votes to, asks, ...etc



- Usage statistics

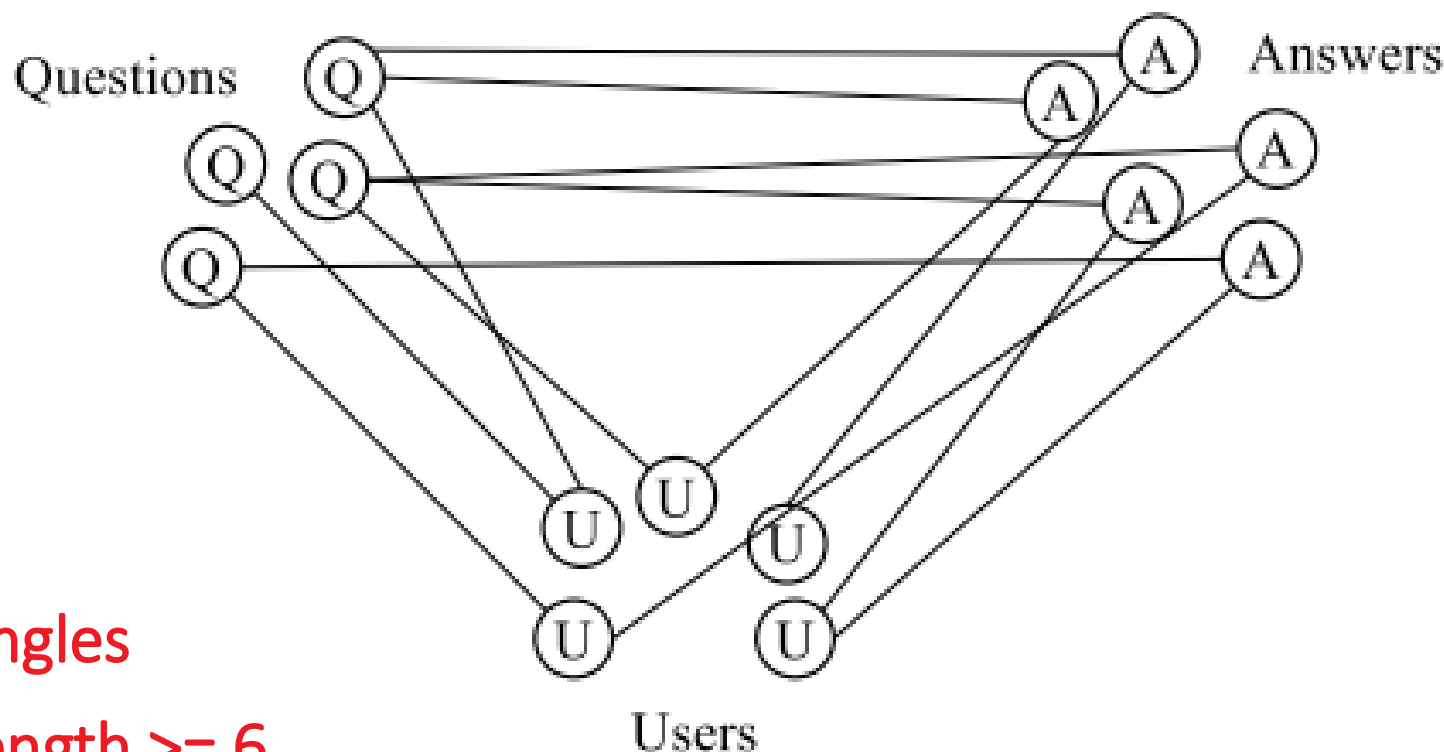
1. Number of clicks
2. Dwell time

- Binary Classification

High quality vs the rest



# Interaction of users-questions-answers

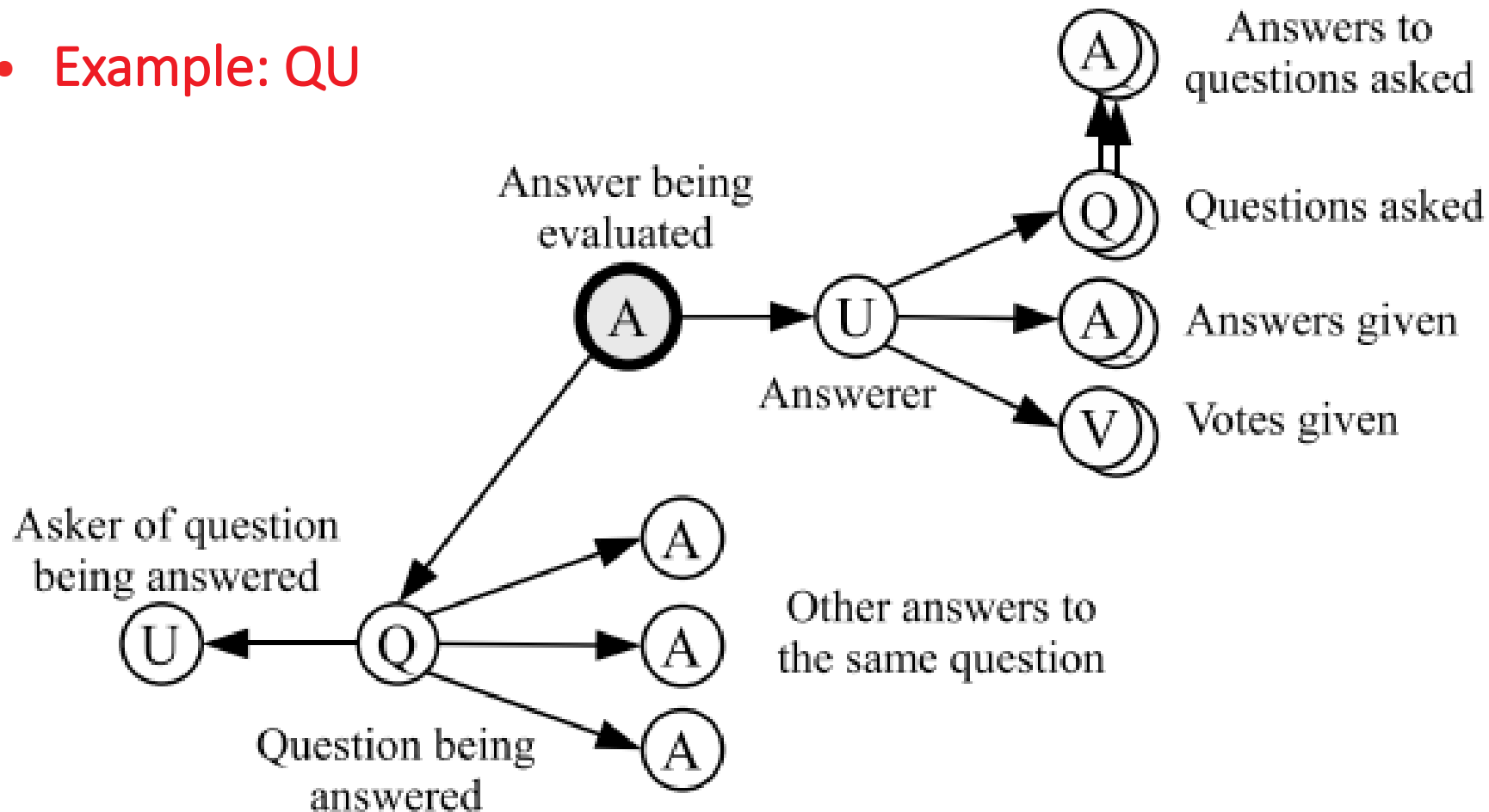


- No triangles
- Cycle length  $\geq 6$

**Figure 2:** Interaction of users-questions-answers modeled as a tri-partite graph.

# Answer Features' Types

- Example: QU



**Figure 3: Types of features available for inferring the quality of an answer.**

# Question Features' Types

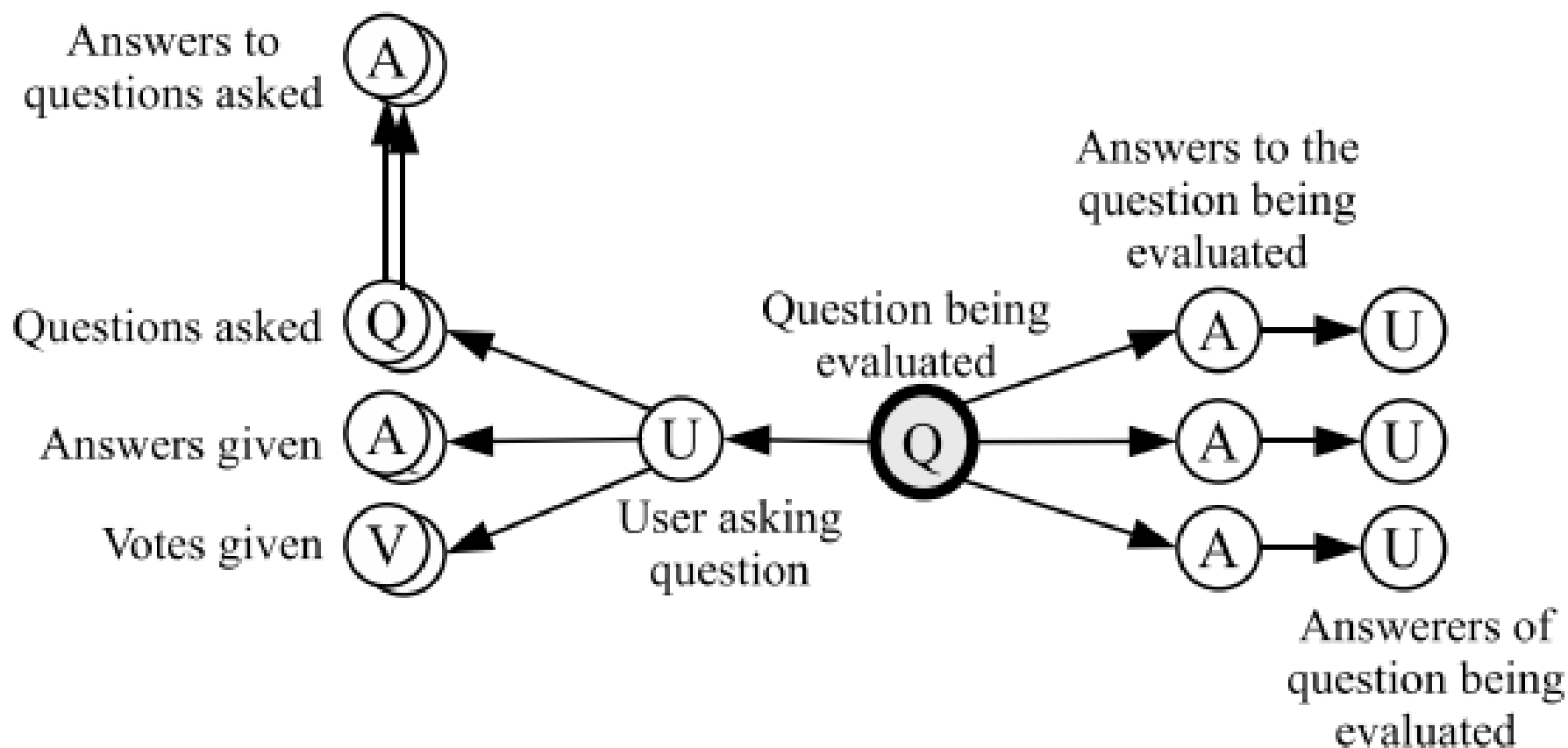


Figure 4: Types of features available for inferring the quality of a question.

- Example: AU

# Implicit User-User Relationships

- Represented by the graph  $G = (V, E)$

Vertices  $V$  = users

Edges  $E = \{ E_a \cup E_b \cup E_v \cup E_s \cup E_+ \cup E_- \}$  represents relationships between users  $u$  and  $v$ :

$E_a$  represents the answers:  $(u, v) \in E_a$  iff user  $u$  has answered at least one question asked by user  $v$

$E_b$  represents the best answers:  $(u, v) \in E_b$  iff user  $u$  has provided at least one best answer to a question asked by user  $v$ .

$E_v$  represents the votes for best answer:  $(u, v) \in E_v$  iff user  $u$  has voted for best answer at least one answer given by user  $v$ .

$E_s$  represents the stars given to questions:  $(u, v) \in E_s$  iff user  $u$  has given a star to at least one question asked by user  $v$ .

$E_{+/-}$  represents the thumbs up/down:  $(u, v) \in E_{+/-}$  iff user  $u$  has given a “thumbs up/down” to an answer by user  $v$ .

# Experimental Setting

- Dataset:

6,665 Questions

8,366 Questions-Answer pairs

- Labeled for quality by human editors for:

- ✓ Well-formedness

- ✓ Readability

- ✓ Utility

- ✓ Interestingness

- ✓ Answers labeled for correctness

- ✓ Questions are assigned a high-level type: informational, advice, poll, etc

# Good questions get good answers

Table 1: Relationship between question quality and answer quality

Answer Quality	Question Quality		
	A. High	B. Medium	C. Low
A. High	41%	15%	8%
B. Medium	53%	76%	74%
C. Low	6%	9%	18%
Total	100%	100%	100%

**Good answers are written in response to good questions**

**Bad questions are the ones that attract bad answers**

# Question/Answers per user

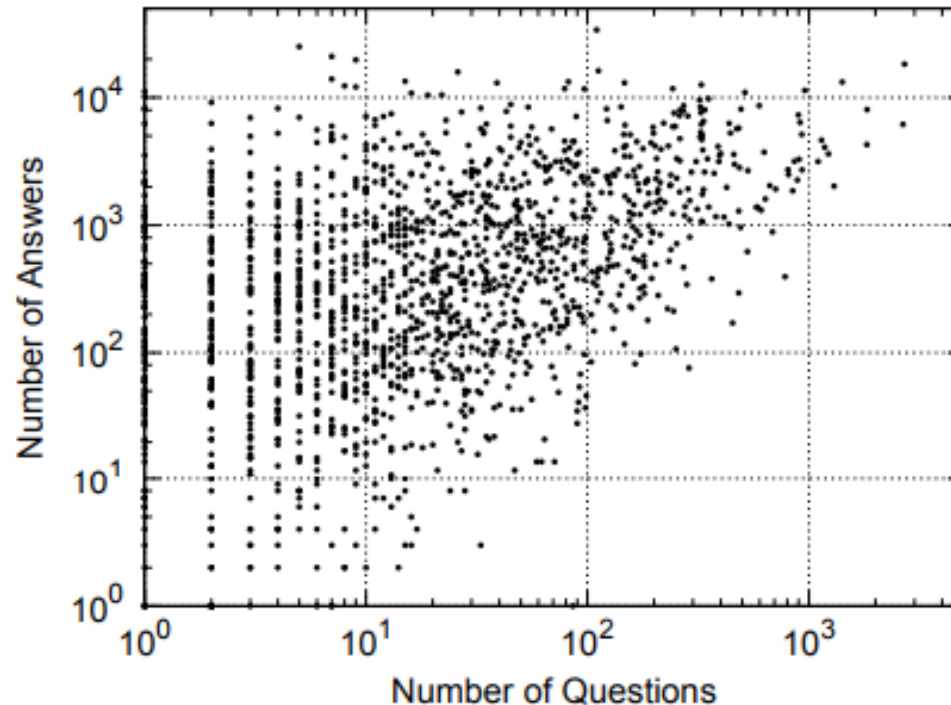


Figure 7: Number of questions and number of answers for each user in our data.

**Heatmap?**

**No clear roles of “asker” and “answerer”**

**Only users with many questions also have many answers is a by-product of the incentive mechanism of the system (points)**

# Question Quality Identification Results

**Table 2: Precision P, Recall R, and Area Under the ROC Curve for the task of finding high-quality questions**

Method	High qual.		Normal/low qual.		AUC
	P	R	P	R	
Text (Baseline)	0.654	0.481	0.762	0.867	0.523
Usage	0.594	0.470	0.755	0.836	0.508
Relation	0.694	0.603	0.806	0.861	0.614
Intrinsic	0.746	0.650	0.829	0.885	0.645
T+Usage	0.683	0.571	0.798	0.865	0.575
T+Relation	0.739	0.647	0.828	0.881	0.659
T+Intrinsic	0.757	0.650	0.830	0.891	0.648
T+Intr.+Usage	0.717	0.690	0.845	0.861	0.686
T+Relation+Usage	0.722	0.690	0.845	0.865	0.679
T+Intr.+Relation	<b>0.798</b>	0.752	0.874	<b>0.901</b>	0.749
All	0.794	<b>0.771</b>	<b>0.885</b>	0.898	<b>0.761</b>



# Finding High Quality Answers Measures

Table 4: Precision P, Recall R, and Area Under the ROC Curve for the task of finding high-quality answers

Method	High qual.		Normal/low qual.		AUC
	P	R	P	R	
Text (Baseline)	0.668	0.862	0.968	0.906	0.805
Relation	0.552	0.617	0.914	0.890	0.623
Intrinsic	0.712	0.918	0.981	0.918	0.869
T+Relation	0.688	0.851	0.965	0.915	0.821
T+Intrinsic	0.711	<b>0.926</b>	<b>0.982</b>	0.917	<b>0.878</b>
All	<b>0.730</b>	0.911	0.979	<b>0.926</b>	0.873

# ROC Curve for Best Performing Classifier

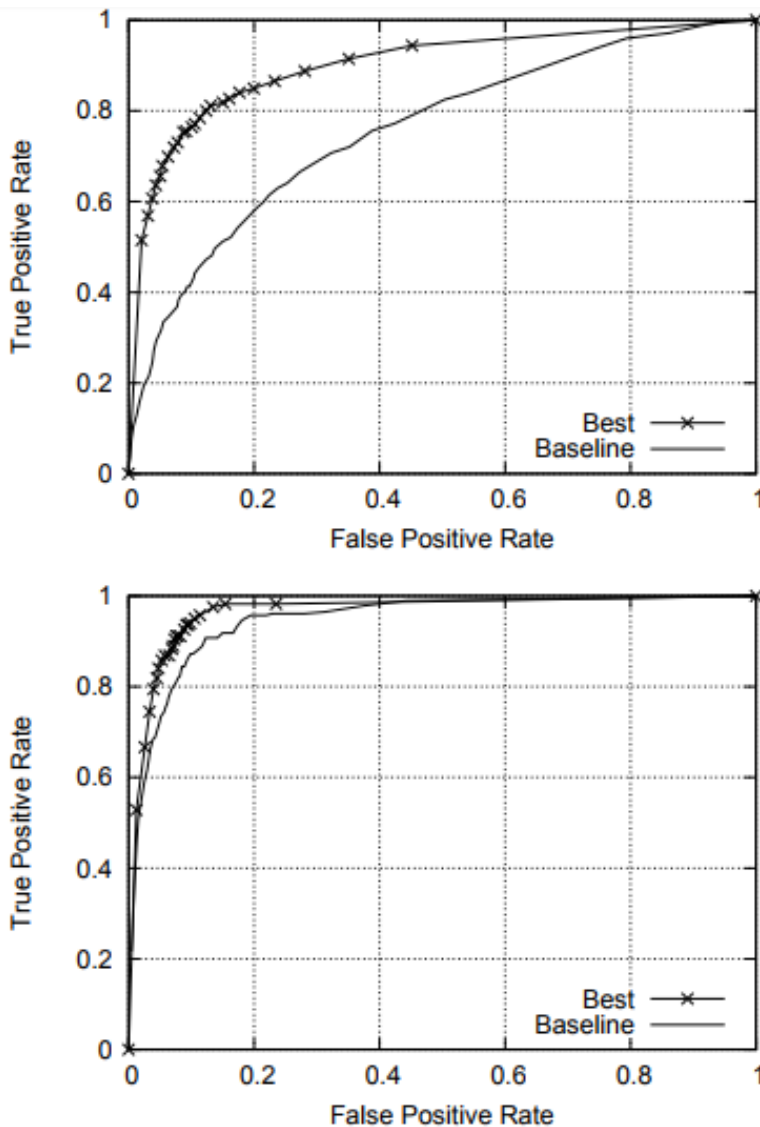


Figure 8: ROC curve for the best-performing classifier, for the task of finding high-quality questions (top) and high-quality answers (bottom).



# Contributions of the paper:

- The paper investigates methods for exploiting community feedback to automatically identify high quality content.
- Develops a comprehensive graph-based model of user relationships, content, and usage features.

## The Second Paper:

- How opinions are received by online communities: a case study on amazon.com helpfulness votes

WWW '09 Proceedings of the 18th international conference  
on World wide web  
Pages 141-150

<https://www.cs.cornell.edu/home/kleinber/www09-helpfulness.pdf>

# Study Outline

- Helpfulness votes:

Examining users' evaluation on online reviews

1. Make hypothesis

2. Proving their validity

3. Coming up with a mathematical model that explains these behaviors

# Amazon Product Star Rating

NEW & INTERESTING FINDS ON AMAZON EXPLORE

amazon [Try Prime](#)

Electronics


There's

Departments


Your Amazon.com 12 Days of Deals Gift Cards Registry Sell Help

EN Hello. Sign in Account & Lists

Computers Deals Laptops Tablets Desktops Monitors Computer Accessories

 \$10 & Under with FREE Shipping  Shop now

Electronics > TV and PC Deal of the Day



Razer

Razer Blade Pro Gaming Laptop - 17" 4K Touchscreen Gaming Laptop (i7-7820HK, 32 GB RAM, 1 TB SSD, GTX 1080 8GB GDDR5X VRAM) - VR Ready

★★★★☆ 167 customer reviews | 143 answered questions

List Price: ~~\$4,399.99~~

Deal of the Day: **\$3,799.00** & **FREE Shipping**. [Details](#)

Ends in 13h 00m 01s

You Save: **\$600.99 (14%)**

**Save \$20 on Microsoft Office with PC** 1 Applicable Promotion

**In Stock.**

**Arrives before Christmas.** Choose delivery option in checkout.

**Want it tomorrow, Dec. 13?** Order within **3 hrs 15 mins** and choose **One-Day Shipping** at checkout. [Details](#)

Average star rating

# Review Helpfulness Rating

## Top customer reviews



Reviewer rating for the product



**Not for Business Users. Heed the warnings of myself and others. Look elsewhere or purchase with expectation of issues.**

July 16, 2017

Style: 14 Inch | Color: 1080P | Capacity: 256 GB | **Verified Purchase**

Not for Business Users. Heed the warnings of myself and others. Look elsewhere or purchase with expectation of issues.

I should have listened to the stories but since I did not heed the warning of others, here is my review to perhaps to provide warning or help you with your purchasing decision.

I purchased this system for one reason, form factor with capability of occasionally gaming. I purchased this laptop primarily for business use because it had a TPM which allows me to take advantage of Bitlocker in Windows to encrypt my drive. I occasionally travel internationally and want the ability to game and not lug around my gaming laptop for the trip.

I read the reviews as you are now and took note of the myriad of reviews indicating multiple issues:

- Critical Hardware Failures within First Year
- Poor Support
- Screen Issues

Helpfulness rating (used to be ratio)

Some people say the laptop is great and I often found their reviews were posted very soon after purchase. The more seasoned reviews painted a

[Read more](#)



[Comment](#)

317 people found this helpful. Was this review helpful to you?

[Report abuse](#)

# Product Rating vs Review Rating

- Product Rating: Opinion

What did **Y** think of **X**?

- Review Helpfulness Rating: Opinion<sup>2</sup>

What did **Z** think of **Y**'s opinion of **X**?

**X**: Product

**Y**: User

**Z**: User



# Hypotheses: Social Mechanisms

- Well-studied hypotheses for how social effects influence group's reaction to an opinion
- The conformity hypothesis
- The individual-bias hypothesis
- The brilliant-but-cruel hypothesis
- The quality-only straw-man hypothesis

# The Conformity Hypothesis

- Review is evaluated as more helpful when its star rating is closer to the consensus star rating
- Helpfulness ratio will be the highest of which reviews have star rating equal to the overall average

# The Individual-Bias Hypothesis

- When a user considers a review, he or she will rate it more highly if it expresses an opinion that he or she agrees with

# The Brilliant-But-Cruel Hypothesis

- Negative reviewers are perceived as more intelligent, competent, and expert than positive reviewers

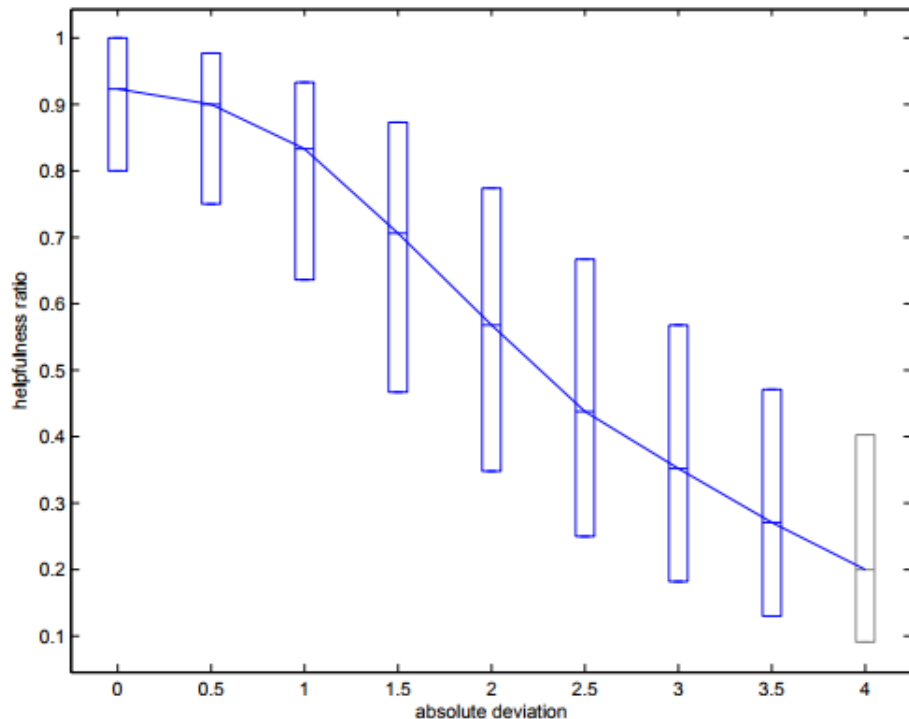
# The Quality-only Straw-man Hypothesis

- Helpfulness is being evaluated purely based on the textual content of reviews
- Non-textual factors are simply correlates of textual quality

# Verifying Hypotheses

- Absolute deviation of helpfulness ratio
- Signed deviation of helpfulness ratio
- Variance of star rating and helpfulness ratio
- Making use of plagiarism

# Absolute Deviation from Average



**Figure 1: Helpfulness ratio declines with the absolute value of a review's deviation from the computed star average; this behavior is predicted by the conformity hypothesis but not ruled out by the other hypotheses.**

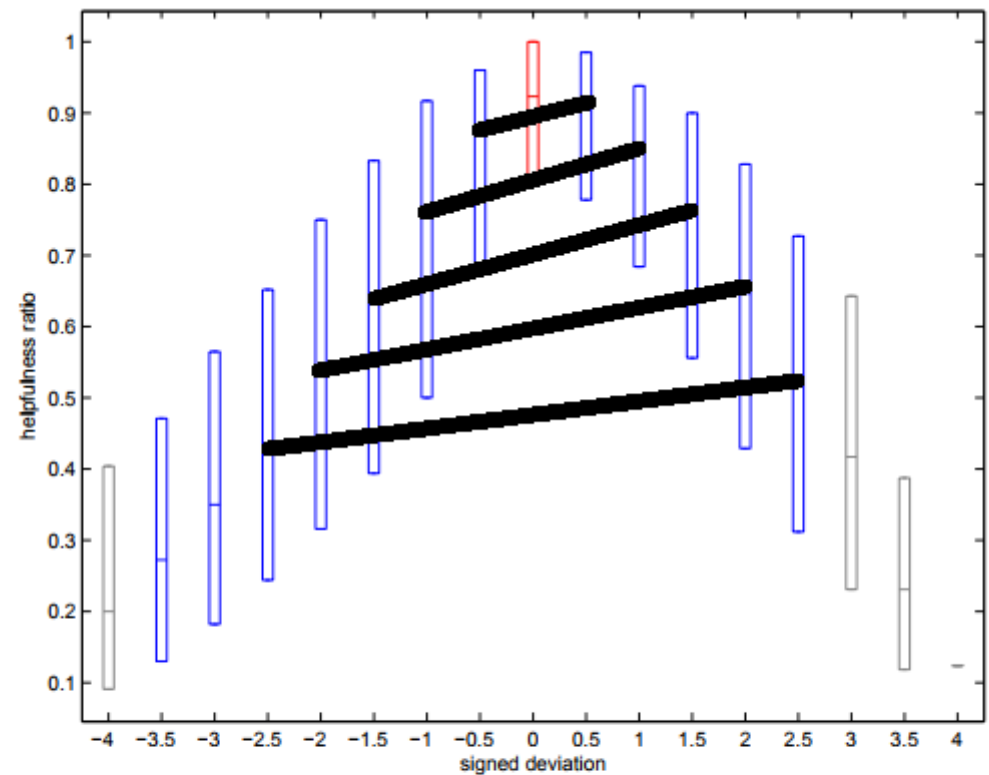
The line segments within the bars (connected by the descending line) indicate the median helpfulness ratio; the bars depict the helpfulness ratio's second and third quantiles.

Throughout, grey bars indicate that the amount of data at that  $x$  value represents .1% or less of the data depicted in the plot.

- Consistent with Conformity hypothesis
- Reviews with star rating close to the average get higher helpfulness ratio

# Signed Deviation from Average

- Not consistent with brilliant-but-cruel hypothesis
- There is tendency towards positivity
- Black lines should not be sloped that way if it is a valid hypothesis



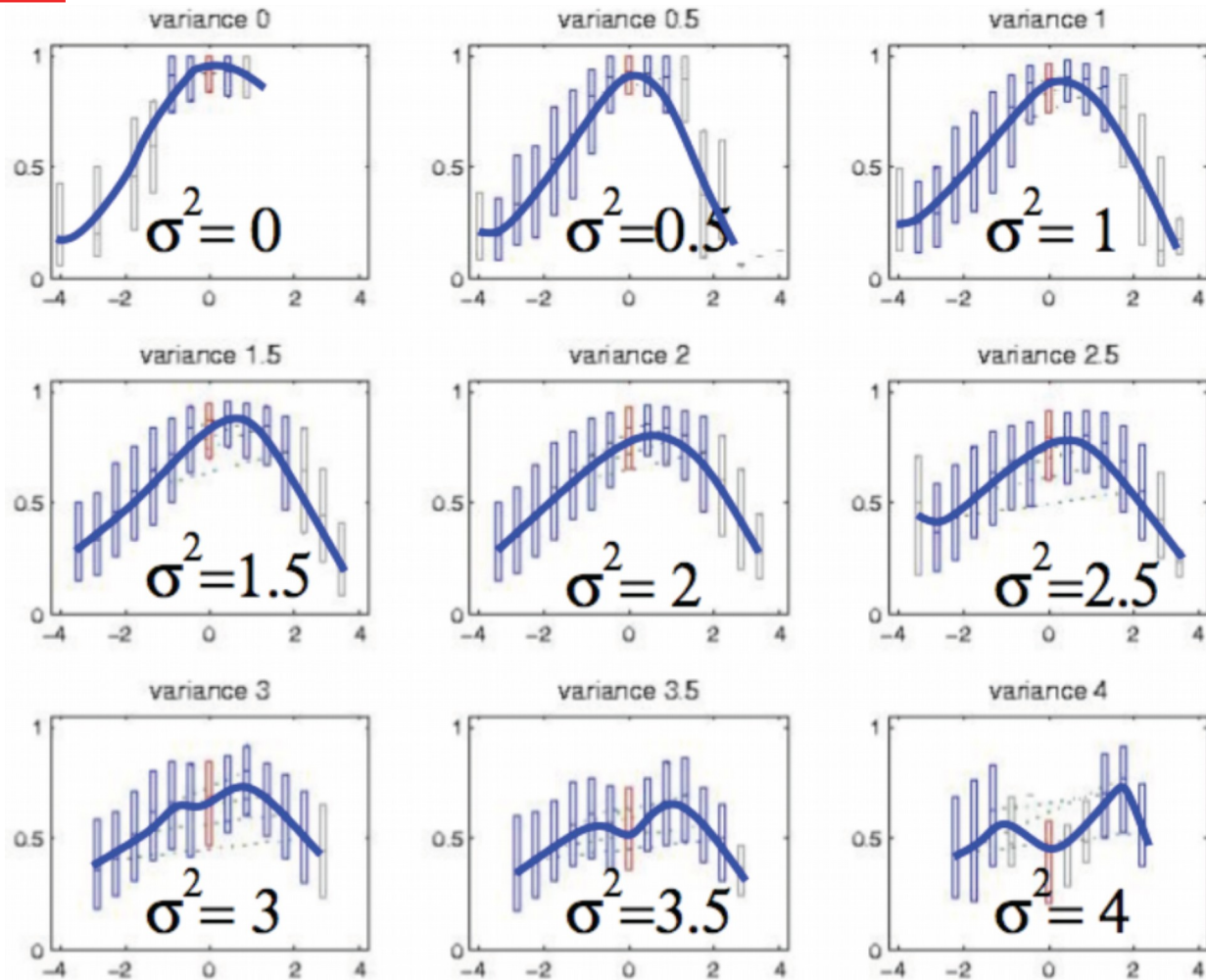
**Figure 2:** The dependence of helpfulness ratio on a review's signed deviation from average is inconsistent with both the brilliant-but-cruel and, because of the asymmetry, the conformity hypothesis.



# Addressing Individual-Bias Effects

- It is hard to distinguish between the conformity and the individual-bias hypothesis
- We need to examine cases in which individual people's opinions do not come from exactly the same distribution
- Cases in which there is high variance in star ratings
- Otherwise conformity and individual-bias are indistinguishable
- Everyone has the same opinion

# Addressing Individual-Bias Effects



**Helpfulness ratio is the highest when star ratings of reviews have average value**

**Helpfulness ratio is the highest with reviews of which rating is slightly-above the average**

**Two-humped camel plots: local minimum around average**

**Figure 3:** As the variance of the star ratings of reviews for a particular product increases, the median helpfulness ratio curve becomes two-humped and the helpfulness ratio at signed deviation 0 (indicated in red) no longer represents the unique global maximum. There are non-zero signed deviations in the plot for variance 0 because we rounded variance values to the nearest .5 increment.

# The Quality-only Straw-man Hypothesis

- Helpfulness is purely based on the textual content of reviews
- Using plagiarized reviews is helpful!
- Challenges:
  1. Almost (no exact) same text
  2. Same exact text could be considered as spam reviews
- If the quality-only straw-man hypothesis holds, helpfulness ratio of “same text” reviews should be the same

# Plagiarism

- Using plagiarism is an effective way to control for the effect of review text
- Definition of plagiarized pair(s) of reviews:  
Two or more reviews of different products with near-complete textual overlap
- Authors consider %70 of textual overlap as plagiarism.

# Plagiarism Example

26 of 30 people found the following review helpful:

★★★★★ **Skull-splitting headache guaranteed!!**, June 16, 2004

By **A Customer**

If you enjoy a thumping, skull splitting migraine headache, then Sing N Learn is for you.

As a longtime language instructor, I agree with the attempt and effort that this series makes, but it is the execution that ultimately weakens Sing N Learn Chinese.

To be sure, there are much, much better ways to learn Chinese. In fact, I would recommend this title only as a last resort and after you've thoroughly exhausted traditional ways to learn Chinese. . . .

7 of 11 people found the following review helpful:

★★★★★ **Migraine Headache at No Extra Charge**, May 28, 2004

By **A Customer**

If you enjoy a thumping, skull splitting migraine headache, then the Sing N Learn series is for you.

As a longtime language instructor, I agree with the effort that this series makes, but it is the execution that ultimately weakens Sing N Learn series. To be sure, there are much, much better ways to learn a foreign language. In fact, I would recommend this title only as a last resort and after you've thoroughly exhausted traditional ways to learn Korean. . . .

**Figure 4: The first paragraphs of “plagiarized” reviews posted for the products Sing 'n Learn Chinese and Sing 'n Learn Korean. In the second review, the title is different and the word “chinese” has been replaced by “korean” throughout. Sources: [http://www.amazon.com/review/RHE2G1M8V0H9N/ref=cm\\_cr\\_rdp\\_perm](http://www.amazon.com/review/RHE2G1M8V0H9N/ref=cm_cr_rdp_perm) and [http://www.amazon.com/review/RQYHTSDUNM732/ref=cm\\_cr\\_rdp\\_perm](http://www.amazon.com/review/RQYHTSDUNM732/ref=cm_cr_rdp_perm).**

# Experiments with Plagiarism

- Text quality is not the only explanatory factor
- Statistically significant difference between the helpfulness ratios of plagiarized pairs

$i \backslash j$	0.5	1	1.5	2	2.5	3	3.5
0		$\succ$	$\succ$	$\succ$	$\succ$	$\succ$	$\succ$
0.5		$\succ$	$\succ$	$\succ$	$\succ$	$\succ$	$\succ$
1			$\succ$	$\succ$	$\succ$	$\succ$	$\succ$
1.5				$\succ$	$\succ$	$\succ$	$\succ$
2					$\succ$	$\succ$	$\succ$
2.5						$\succ$	$\succ$
3							

The plagiarized reviews with deviation 1 is significantly more helpful than those with deviation 1.5

**Table 1:** “Plagiarized” reviews with a lower absolute deviation tend to have larger helpfulness ratios than duplicates with higher absolute deviations. Depicted: whether reviews with deviation  $i$  have an helpfulness ratio significantly larger ( $\succ$ ) or significantly smaller ( $\prec$ , no such cases) than duplicates with absolute deviation  $j$  (blank: no significant difference).

# What to Take Away?

- A review's perceived helpfulness depends not just on its content, but also the relation of its score to other scores
- The dependence of the score is consistent with a simple and natural model of individual-bias in the presence of a mixture of opinion distributions



# References:

- [Finding high-quality content in social media](#)  
WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining  
Pages 183-194  
<http://www.mathcs.emory.edu/~eugene/papers/wsdm2008quality.pdf>
- [How opinions are received by online communities: a case study on amazon.com helpfulness votes](#)  
WWW '09 Proceedings of the 18th international conference on World wide web  
Pages 141-150  
<https://www.cs.cornell.edu/home/kleinber/www09-helpfulness.pdf>



## Why did the 2<sup>nd</sup> paper cite the 1<sup>st</sup> paper?

- The second paper studied elements that affect the helpfulness/quality of reviews which took advantage of some of the methods studied in the first paper for identifying high quality content on social media.