Bots, Socks, and Vandals: Malicious Actors on the Web

V.S. Subrahmanian
Dartmouth College
vs@dartmouth.edu
@vssubrah

Joint work with many students, postdocs, and colleagues !

@vssubrah
vs@dartmouth.edu

How Trolls Are Ruining the Internet

Wikipedia blocks hundreds of 'scam' sock puppet accounts

When Will the Internet Be Safe for Women?

FAKE NEWS IS ABOUT TO GET EVEN SCARIER THAN YOU EVER DREAMED

Fake reviews on the Play Store reportedly growing and getting smarter

Time (2016); The Atlantic (2016); BBC (2015), Vanity Fair (2017), Digital Trends (20

@vssubrah
vs@dartmouth.edu

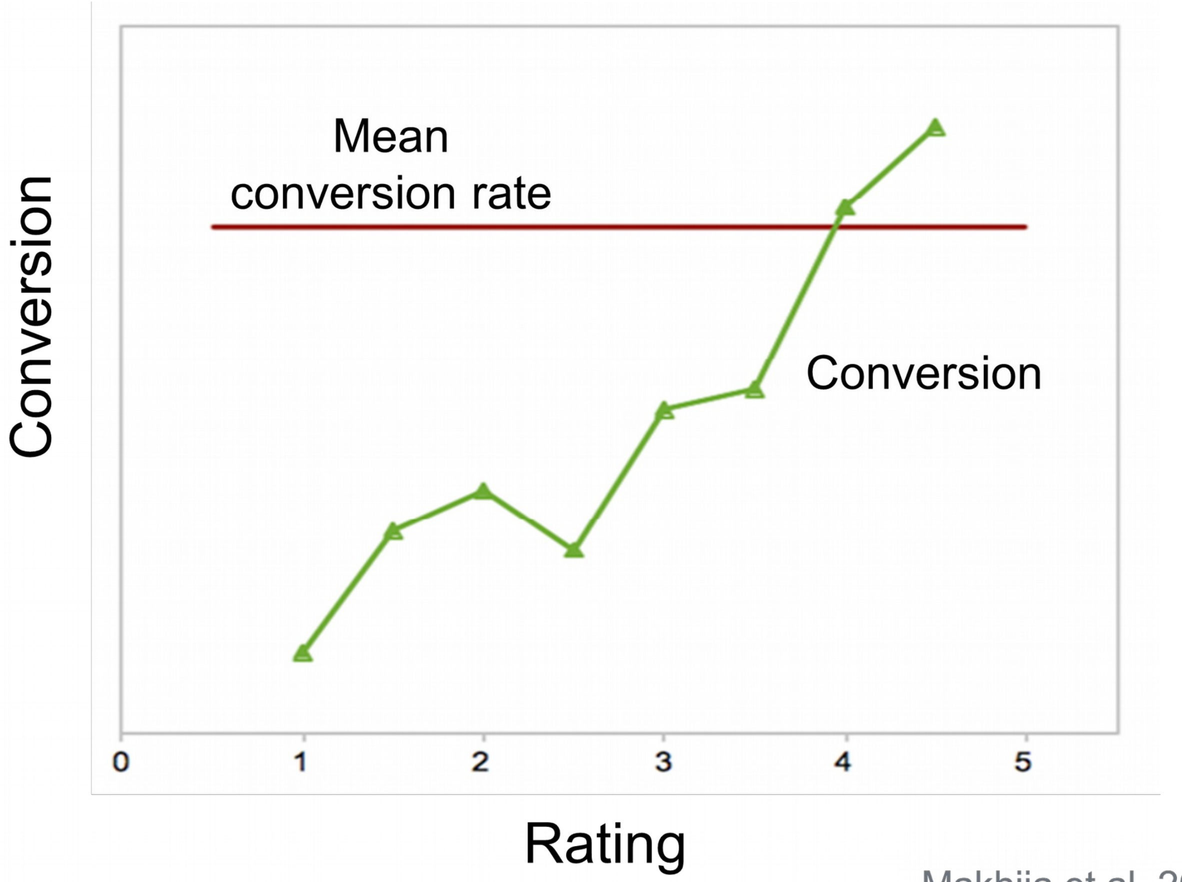
Outline of talk

- **Online Marketplaces: Review Fraud**
- News & Other Discussion Forms: Sockpuppet Accounts
- Wikipedia: Vandals
- Twitter: Bots
- Malicious Actors – The Next Generation

Rev2: Fraudulent User Prediction in Rating Platforms
S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos and V.S. Subrahmanian.
WSDM 2018. **Used in production at Flipkart, India**

@vssubrah
vs@dartmouth.edu

Review Fraud → Increased Revenues

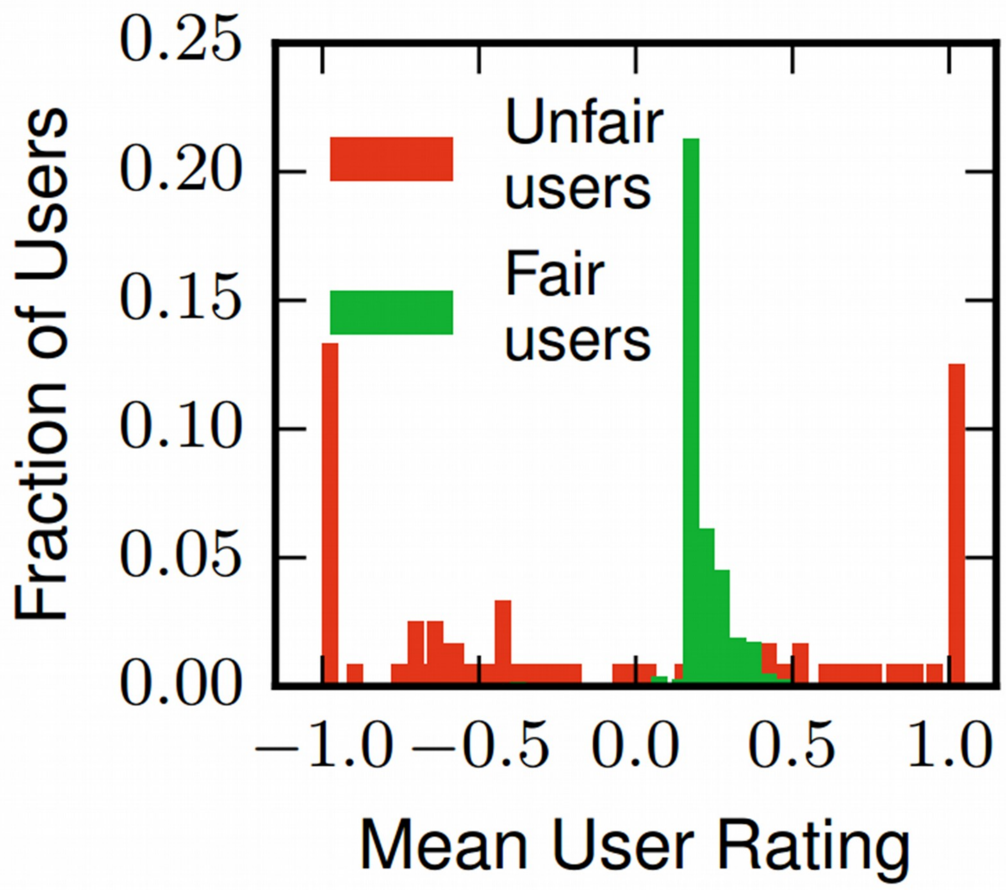


rating increases revenue by 5-9% on Yelp (*Luca et al., Management Sci.,*

Makhija et al, 2016, Luca et al., 2016

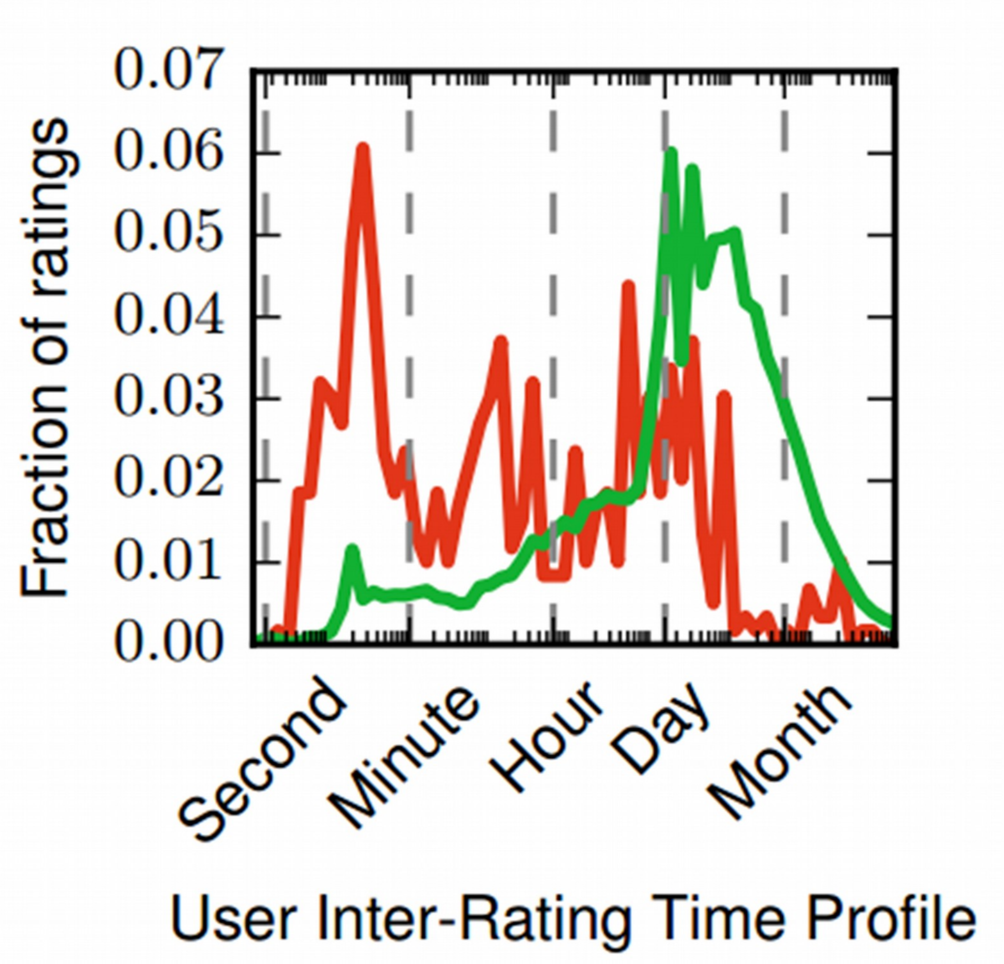
@vssubrah
vs@dartmouth.edu

Review Fraud, I: Review Fraudsters Have Stronger Opinions



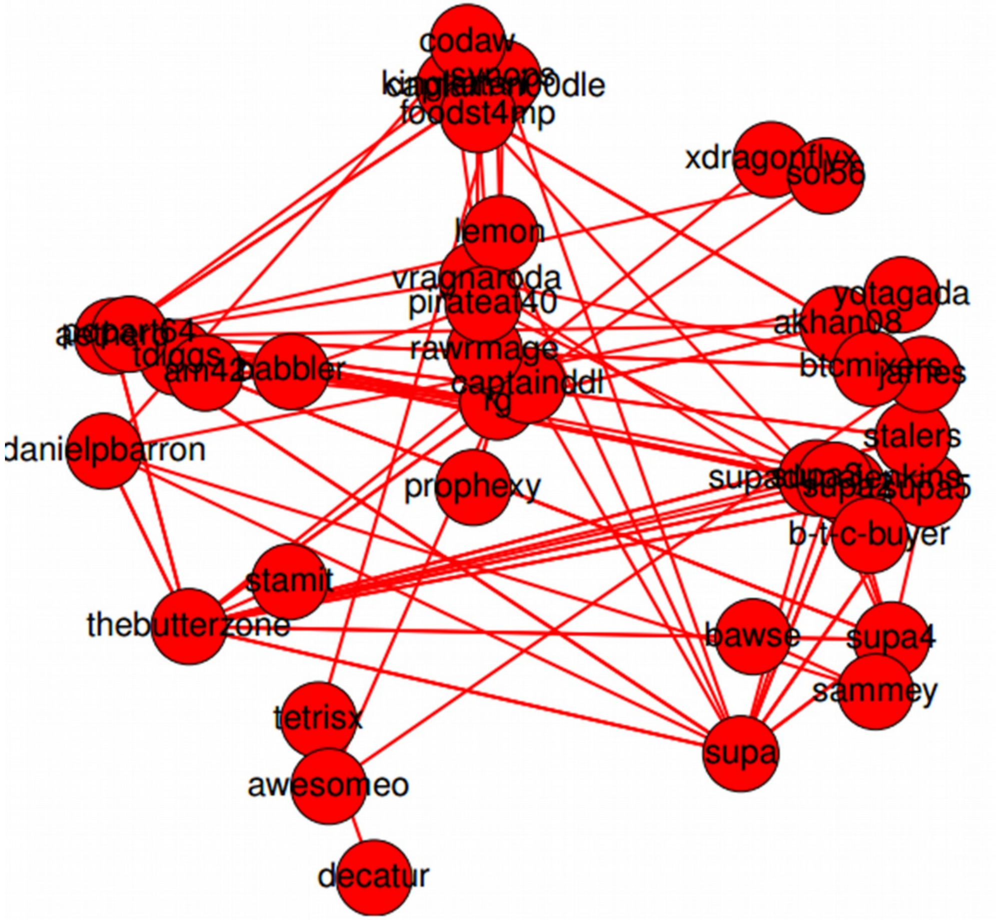
@vssubrah
vs@dartmouth.edu

Review Fraud, II: Review Fraudsters Generate Reviews Faster



@vssubrah
vs@dartmouth.edu

Review Fraud, III: Fraudsters Review Each Other Positively



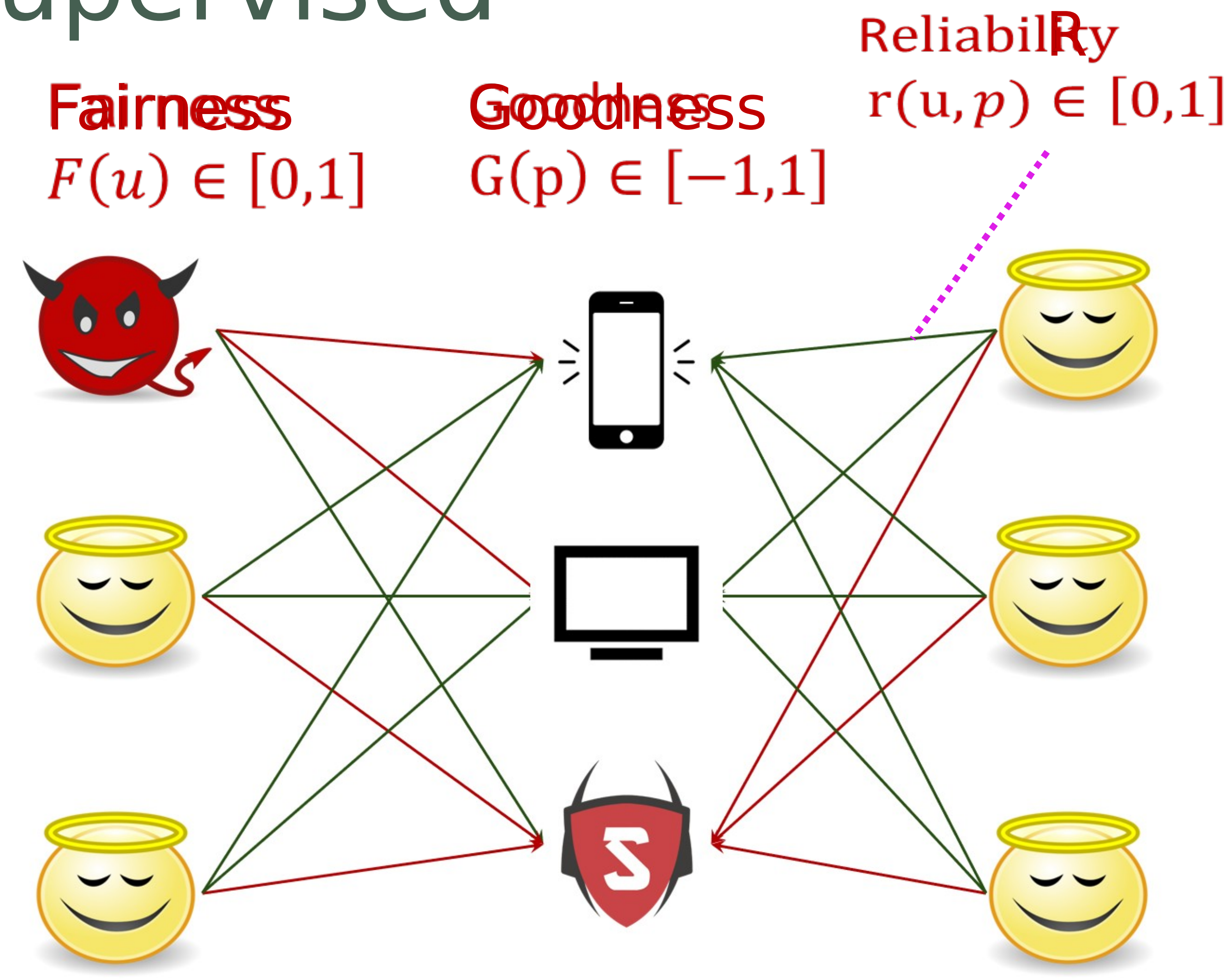
Sample Discoveries on Bitcoin Alpha

- REV2 automatically identified a coordinated group/cluster of users who
  - Rate others in the group positively
  - Rate many outside the group negatively
- Past efforts that use rating and time distributions are unable to identify

@vssubrah
vs@dartmouth.edu

REV2 Algorithm: Unsupervised

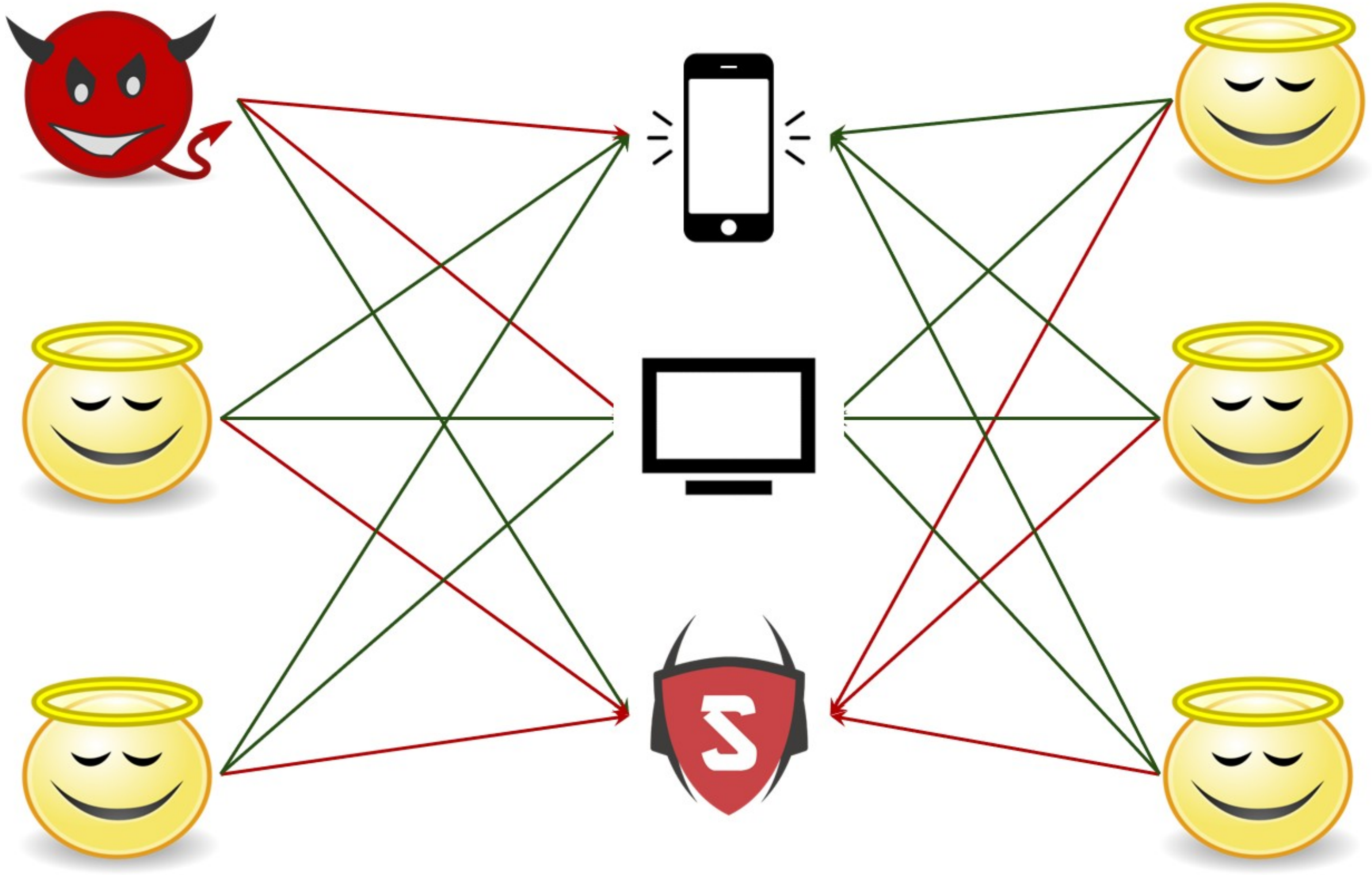
- Represents data via a bipartite graph consisting of three kinds of entities
  - *User* nodes: Authors of reviews.
    - Each user *u* has an associated *fairness f(u).*
  - *Product* nodes: Subject of reviews.
    - Each product *p* has an associated goodness *g(p).*
  - *Review edges:* Link users to products they have reviewed.
    - Each review *r* has an associated reliability *rel(r).*
- **We have to discover these**



Red edge = -1, green edge = +1 rating

@vssubrah
vs@dartmouth.edu

REV2: Fairness

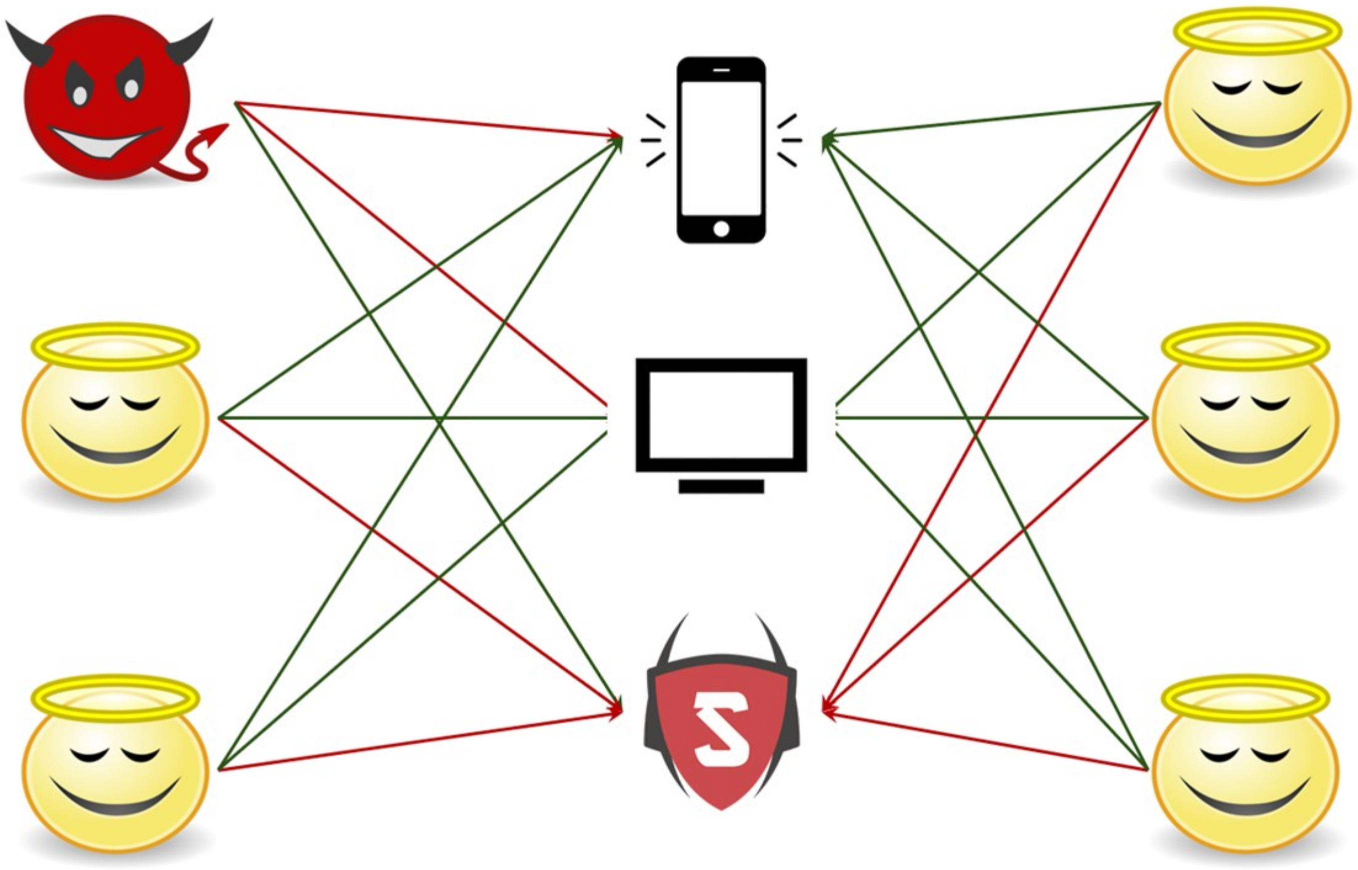


$$F(u) = \frac{\sum_{(u,p) \in \mathrm{Out}(u)} R(u,p)}{|\mathrm{Out}(u)|}$$

Fairness = average reliability of user's reviews.

@vssubrah
vs@dartmouth.edu

REV2: Goodness



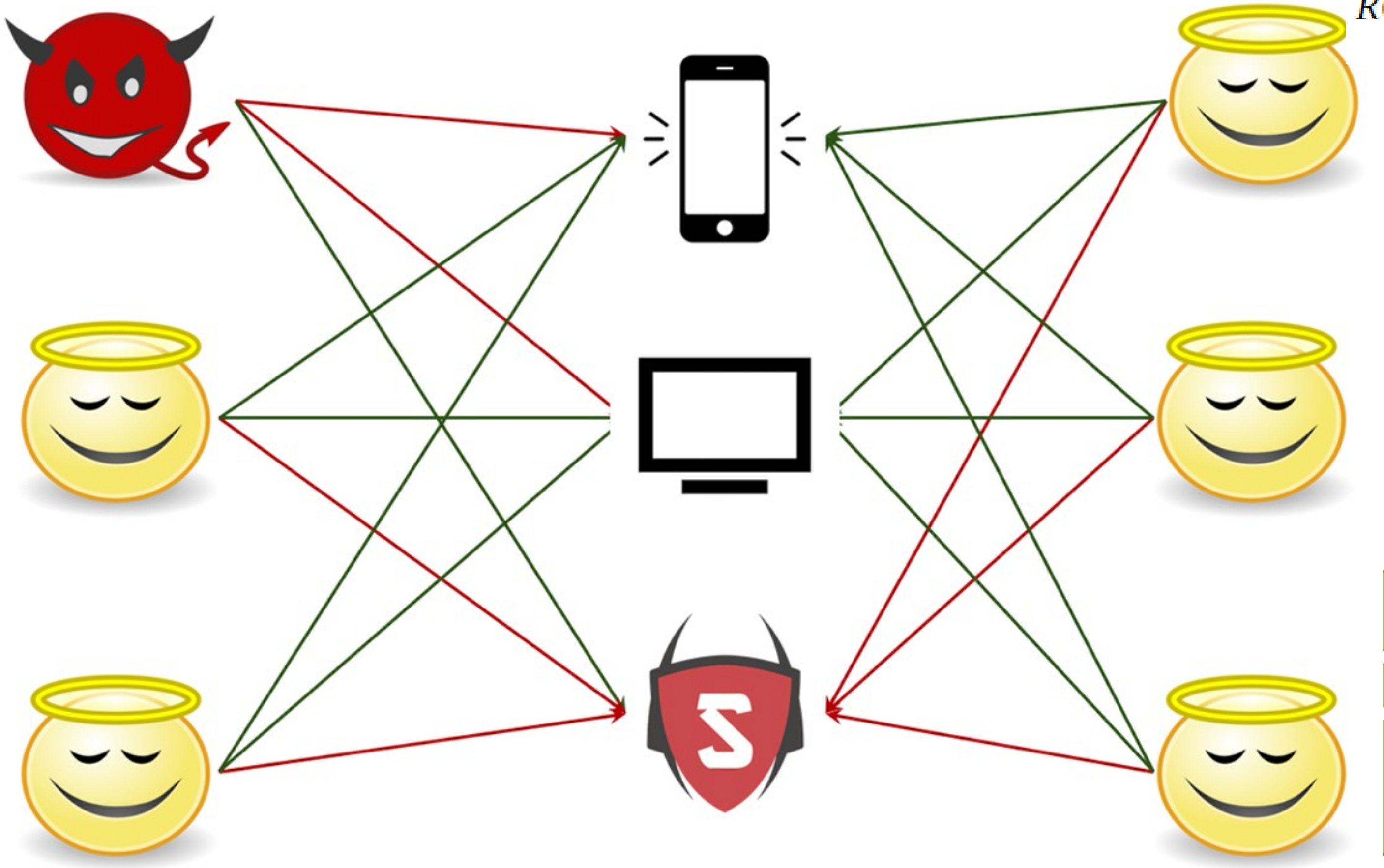
$$G(p) = \frac{\sum_{(u,p) \in \mathtt{In(p)}} R(u,p) \cdot \mathtt{score(u,p)}}{|\mathtt{In(p)}|}$$

discounted rating of a single review.

- Summation: Expected sum of discounted ratings of all reviews of a product.

@vssubrah
vs@dartmouth.edu

REV2: Reliability

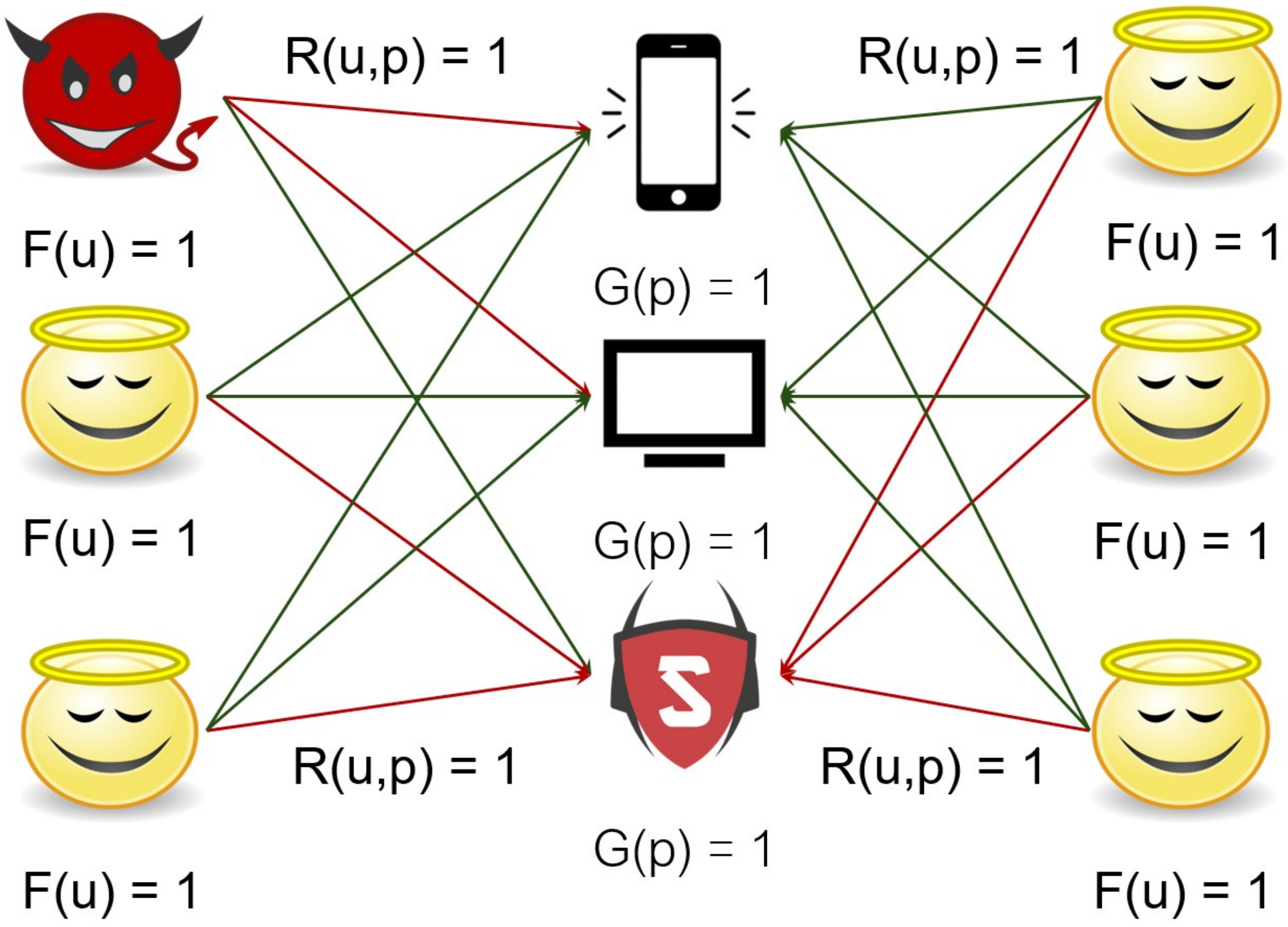


$$R(u,p) = \frac{1}{\gamma_1 + \gamma_2}(\gamma_1 \cdot F(u) + \gamma_2 \cdot (1 - \frac{|\text{score}(u,p) - G(p)|}{2}))$$

- Fairness of the user
- Deviation of user's score from goodness – penalize high deviation

@vssubrah
vs@dartmouth.edu

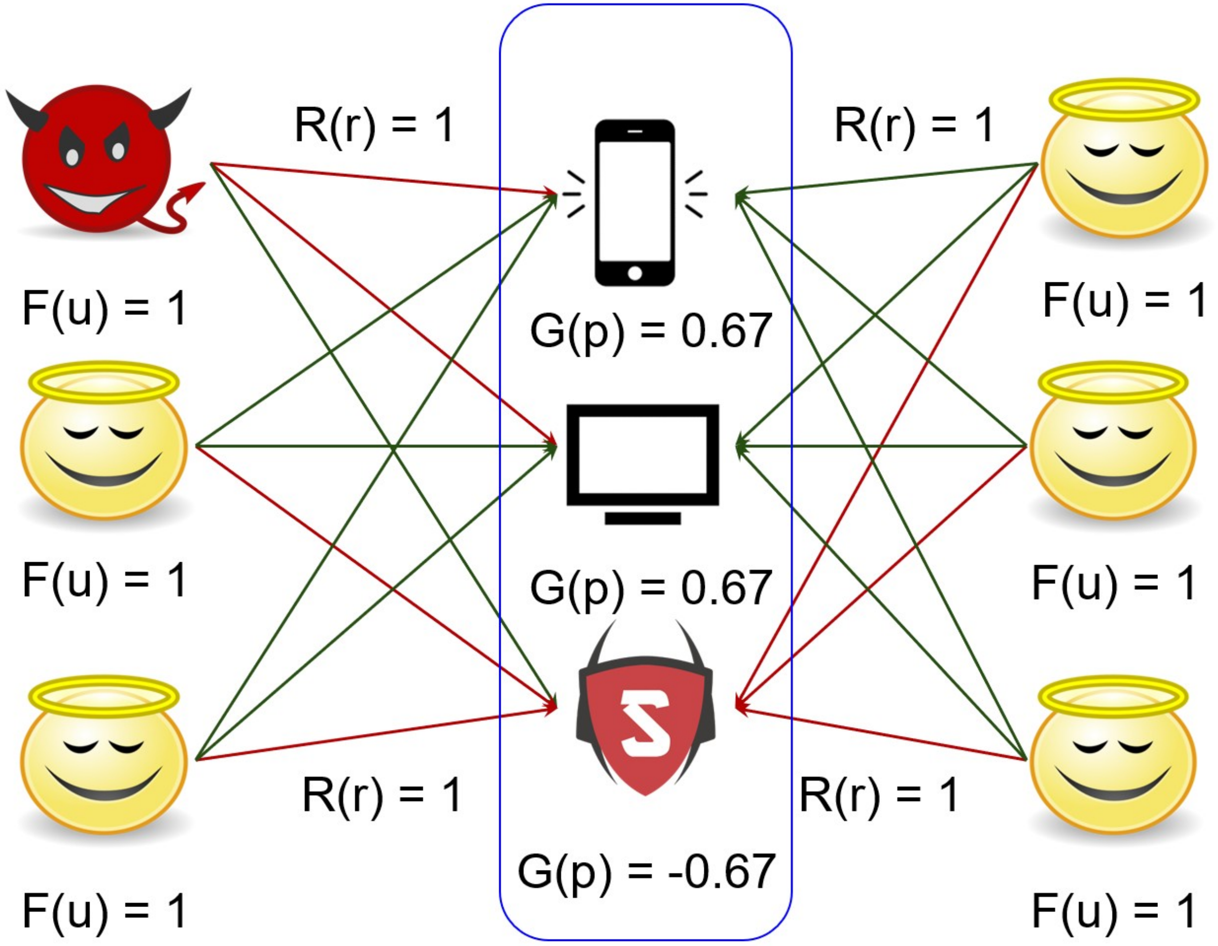
REV2 Algorithm: Initialization



Initialize all variables to 1

@vssubrah
vs@dartmouth.edu

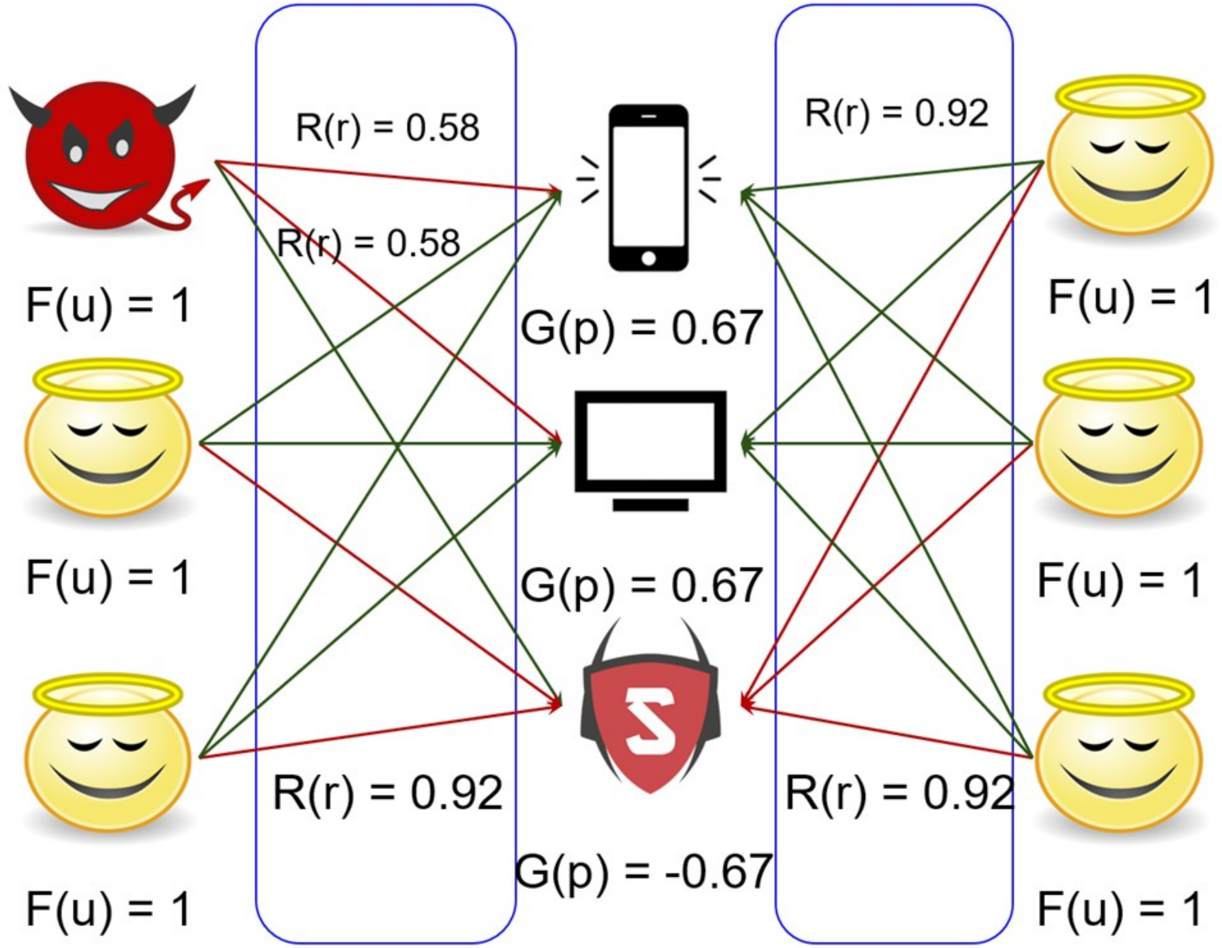
REV2 Algorithm: Update Goodness



$$G(p) = \frac{\sum_{(u,p)\in \mathtt{In(p)}} R(u,p)\cdot \mathtt{score(u,p)}}{|\mathtt{In(p)}|}$$

@vssubrah
vs@dartmouth.edu

REV2 Algorithm: Update Reliability

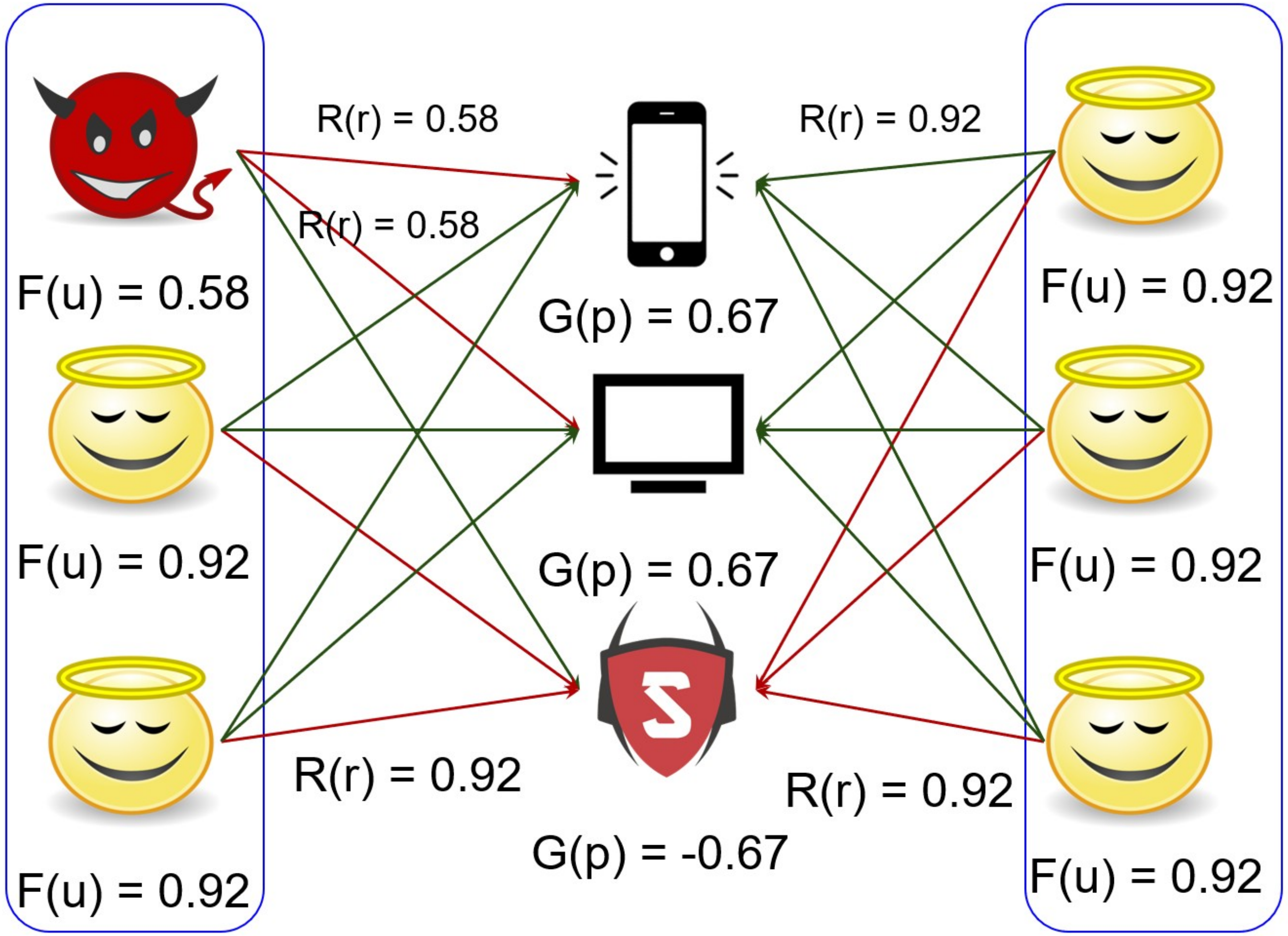


$$R(u,p) = \frac{1}{\gamma_1 + \gamma_2}(\gamma_1 \cdot F(u) + \gamma_2 \cdot (1 - \frac{|\text{score}(u,p) - G(p)|}{2}))$$

Used $\gamma_1 = \gamma_2 = 1$ in the example.

@vssubrah
vs@dartmouth.edu

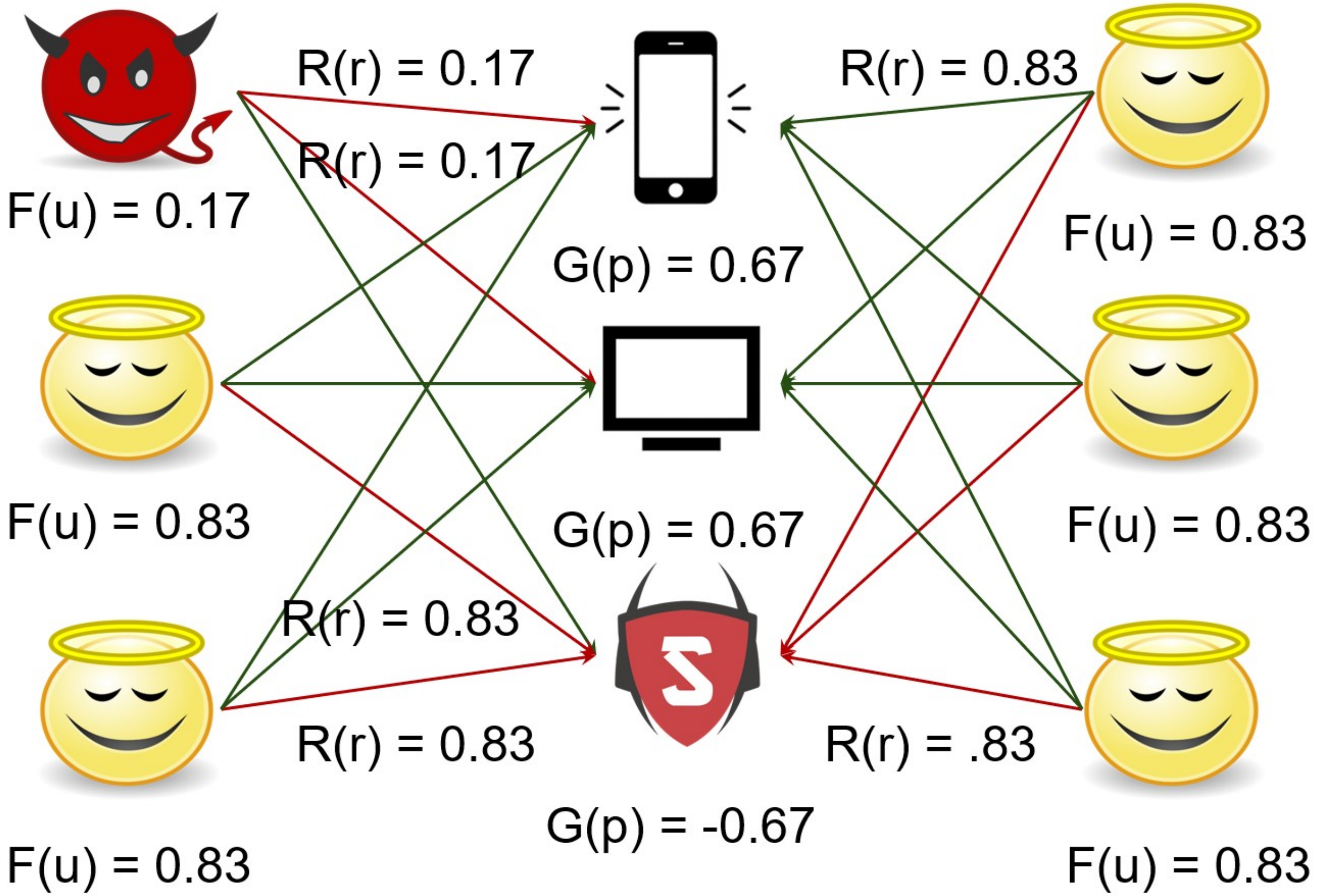
REV2 Algorithm: Update Fairness



$$F(u) = \frac{\sum_{(u,p)\in \mathrm{Out}(u)} R(u,p)}{|\mathrm{Out}(u)|}$$

@vssubrah
vs@dartmouth.edu

REV2 Algorithm: Convergence State



@vssubrah
vs@dartmouth.edu

But…. Cold Start Problem

- Most products get only a few ratings
- Most reviewers provide only a small number of reviews
- Add Bayesian Priors

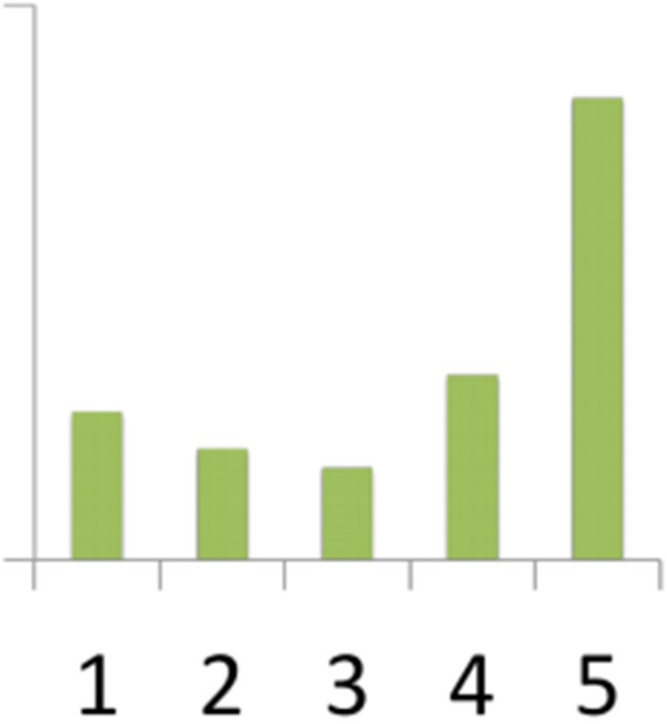
$$F(u) = \frac{\sum_{(u,p)\in \mathrm{Out(u)}} R(u,p) + \alpha_1 \cdot \mu_f}{|\mathrm{Out(u)}| + \alpha_1}$$

$$G(p) = \frac{\sum_{(u,p)\in \mathrm{In(p)}} R(u,p) \cdot \mathrm{score(u,p)} + \beta_1 \cdot \mu_g}{|\mathrm{In(p)}| + \beta_1}$$

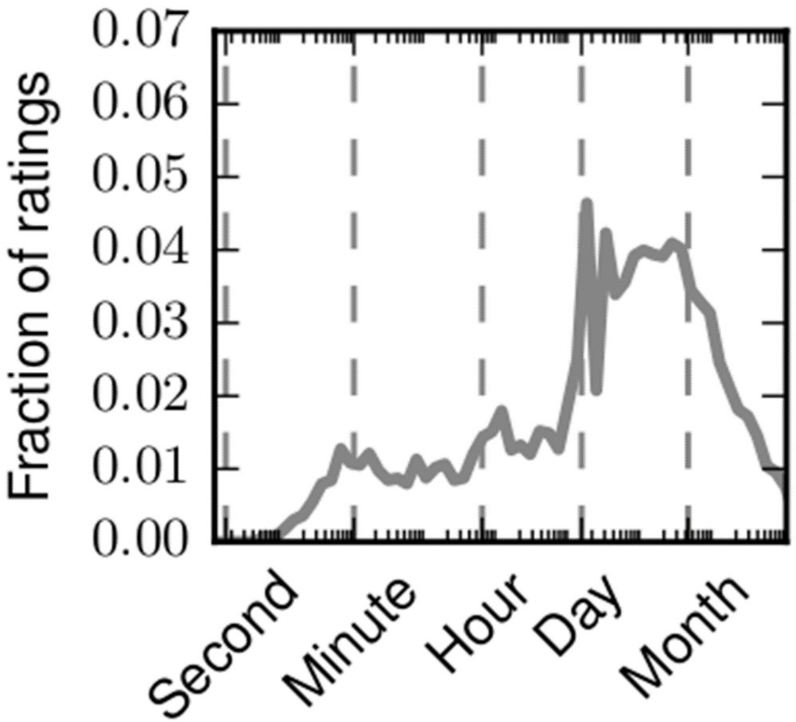
- μ_f, μ_p values are mean fairness and goodness scores over all user and product nodes, respectively.
- $\alpha_1, \beta_1 \geq 0$ are weight denoting importance of the mu values.

@vssubrah
vs@dartmouth.edu

But….: What about Behavioral Properties?



Rating distribution



Time-stamp distribution

Use BIRDNEST score of reviewers and products (Hooi et al., 2016)

@vssubrah
vs@dartmouth.edu

Updated REV2 Formulas

$$F(u) = \frac{\sum_{(u,p)\in \mathrm{Out(u)}} R(u,p) + \alpha_1 \cdot \mu_f + \alpha_2 \cdot \Pi_U(u)}{|\mathrm{Out(u)}| + \alpha_1 + \alpha_2}$$

$$R(u,p) = \frac{\gamma_1 \cdot F(u) + \gamma_2 \cdot \left(1 - \frac{|\mathrm{score(u,p)} - G(p)|}{2}\right) + \gamma_3 \cdot \Pi_R(u,p)}{\gamma_1 + \gamma_2 + \gamma_3}$$

$$G(p) = \frac{\sum_{(u,p)\in \mathrm{In(p)}} R(u,p) \cdot \mathrm{score(u,p)} + \beta_1 \cdot \mu_g + \beta_2 \cdot \Pi_P(p)}{|\mathrm{In(p)}| + \beta_1 + \beta_2}$$

Cold start treatment

Behavioral property scores

@vssubrah
vs@dartmouth.edu

Unsupervised Prediction

	Unfair user prediction					Fair user prediction				
	OTC	Alpha	Amazon	Epinions	Flipkart	OTC	Alpha	Amazon	Epinions	Flipkart
FraudEagle	93.67	86.08	47.21	*nc*	*nc*	86.94	71.99	96.88	*nc*	*nc*
BAD	79.75	63.29	55.92	58.31	79.96	77.41	68.31	97.19	97.09	38.07
SpEagle	74.40	68.42	12.16	*nc*	*nc*	80.91	82.23	93.42	*nc*	*nc*
BIRDNEST	61.89	53.46	19.09	37.08	85.71	46.11	77.18	93.32	98.53	62.47
Trustiness	74.11	49.40	40.05	*nc*	*nc*	84.09	78.19	97.33	*nc*	*nc*
REV2	96.30	75.29	64.89	81.56	99.65	92.85	84.85	100.0	99.81	42.83

@vssubrah
vs@dartmouth.edu

Supervised Prediction (using Random Forest)

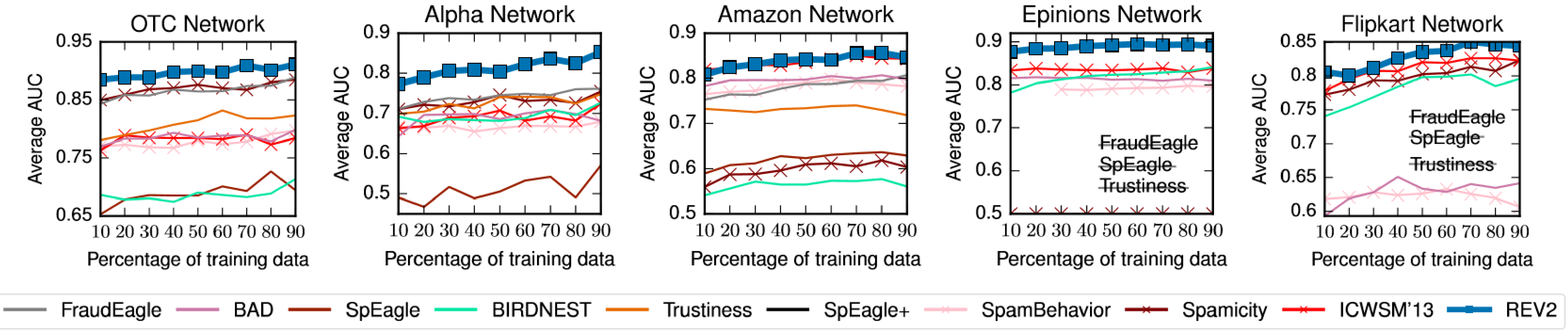
	OTC	Alpha	Amazon	Epinions	Flipkart
FraudEagle	0.89	0.76	0.81	*nc*	*nc*
BAD	0.79	0.68	0.80	0.81	0.64
SpEagle	0.69	0.57	0.63	*nc*	*nc*
BIRDNEST	0.71	0.73	0.56	0.84	0.80
Trustiness	0.82	0.75	0.72	*nc*	*nc*
SpEagle+	0.55	0.66	0.67	*nc*	*nc*
SpamBehavior	0.77	0.69	0.80	0.80	0.60
Spamicity	0.88	0.74	0.60	0.50	0.82
ICWSM'13	0.75	0.71	0.84	0.82	0.82
REV2	0.90	0.88	0.85	0.90	0.87

127 of 150 reported fake reviewers in Flipkart correct.
REV2 is in use at Flipkart.

- 2000+ features
- Combinations of fairness/goodness scores under various parameter settings

@vssubrah
vs@dartmouth.edu

Robustness of REV2



REV2 provides robust predictions regardless of the amount of data used for training.

@vssubrah
vs@dartmouth.edu

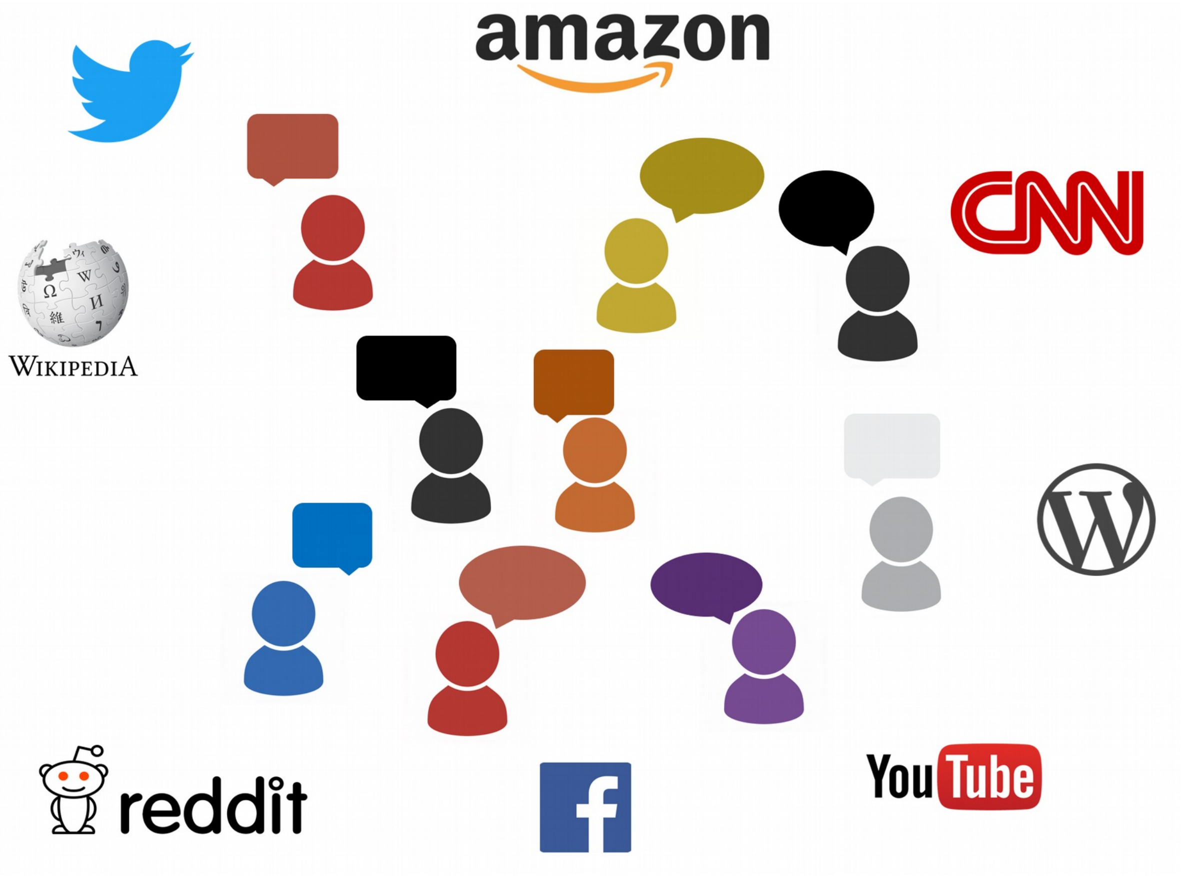
Outline of talk

- Online Marketplaces: Review Fraud
- **News & Other Discussion Forms: Sockpuppet Accounts**
- Wikipedia: Vandals
- Twitter: Bots
- Malicious Actors – The Next Generation

An Army of Me: Sockpuppets in Online Discussion Communities.
S. Kumar, J. Cheng, J. Leskovec and V.S. Subrahmanian. Proceedings of the 26th International World Wide Web Conference (WWW), 2017.
Best Paper Award Honorable Mention
Being transitioned to both Wikipedia and Reddit.

@vssubrah
vs@dartmouth.edu

Sockpuppets



DISQUS

CNN BREITBART FOX NEWS THE HILL allkpop npr A.V. CLUB IGN MLB.com

2.9M Users

2.1M Articles

62M Posts

@vssubrah
vs@dartmouth.edu

Sock Example

Why DC is better than Marvel

IGN

April 28, 2013 by Eric_17

bdiaz209 April 28 2013, 11PM
Possibly the best blog I've ever read major props to you

Eric_17 April 28 2013, 12AM
Thanks. I knew Marvel fans would try to flame me, but they have nothing other than "oh that's your opinion" instead of coming up with their own argument

Fellstrike April 29 2013, 6PM
Quit talking to yourself, *******. Get back on your meds if you're going to do that

bdiaz209 only posts on this discussion to support and defend Eric_17

@vssubrah
vs@dartmouth.edu

Defining Socks

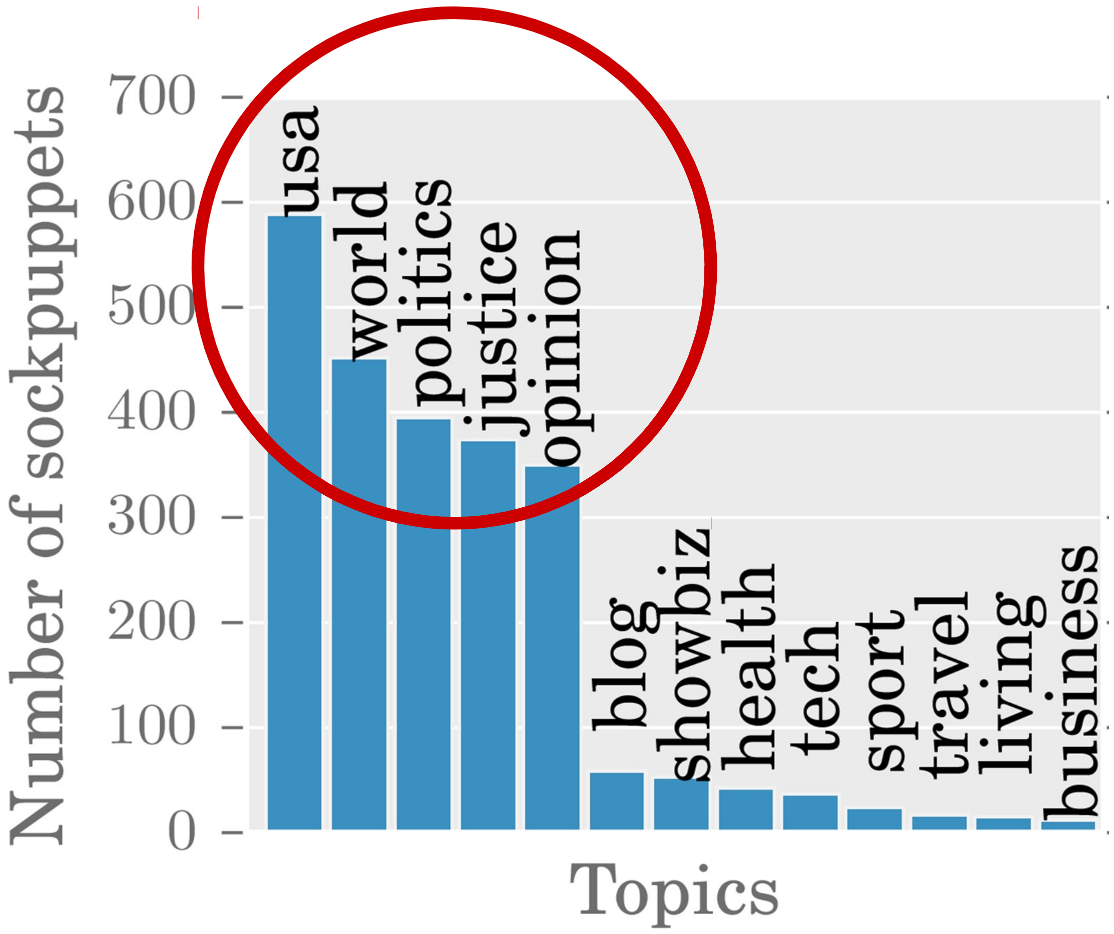
Sockpuppets are accounts that post from the same IP address in the same discussion very close in time (15 min), in at least 3 different instances.

3656 Sockpuppets
1653 puppet masters

IP addresses only used for ground truth, not for prediction.

@vssubrah
vs@dartmouth.edu

Where do Sockpuppet Accounts Post?



@vssubrah
vs@dartmouth.edu

How do sockpuppets write?

jakey008 Feb 5 2013, 2PM
should have read the reviews first :(

ricobeans27 Feb 5 2013, 3PM
Couldn’t agree more!!

Falcon-X32 Feb 5 2013, 3PM
I agree. You are absolutely right!

Agree more
$p < 10^{-3}$

More self centered
$p < 10^{-3}$

Use short sentences
$p < 10^{-3}$

Address others directly
$p < 10^{-3}$

Down-voted more
$p < 10^{-3}$

Start fewer discussions
$p < 10^{-3}$

@vssubrah
vs@dartmouth.edu

How do sockpuppets interact?

jakey008 Feb 5 2013, 2PM
should have read the reviews first :(

ricobeans27 Feb 5 2013, 3PM
Couldn’t agree more.

Upvote each other more
$p < 10^{-3}$

Falcon-X32 Feb 5 2013, 3PM
I agree. You are absolutely right!

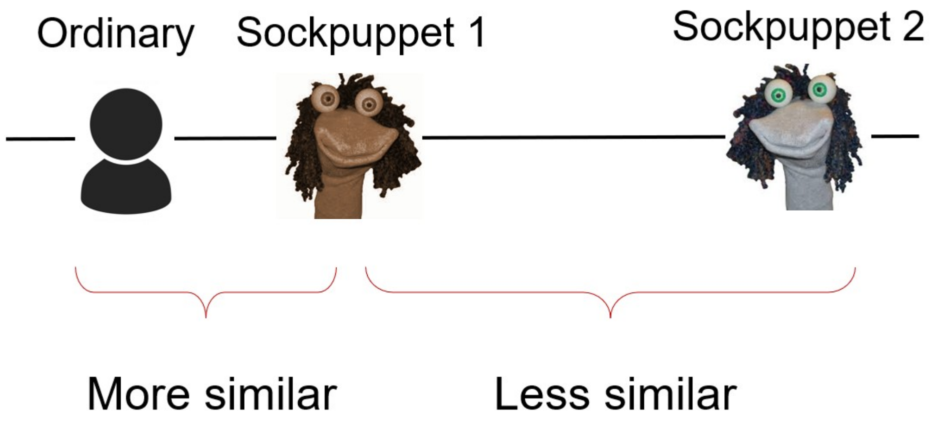
Smoothzilla Feb 5 2013, 3PM
Thanks for your support!!!!

Interact more with each other
$p < 10^{-3}$

@vssubrah
vs@dartmouth.edu

Double-Life Hypothesis

Double life hypothesis:
Puppetmaster maintains distinct personality for the two sockpuppets

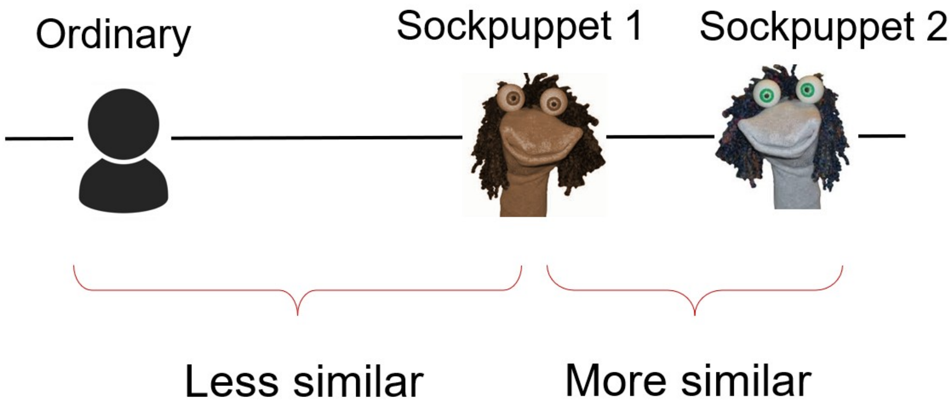


Similarity is measured as cosine similarity between user posts' features: LIWC, sentiment, number of words, etc.

@vssubrah
vs@dartmouth.edu

Alternate Hypothesis

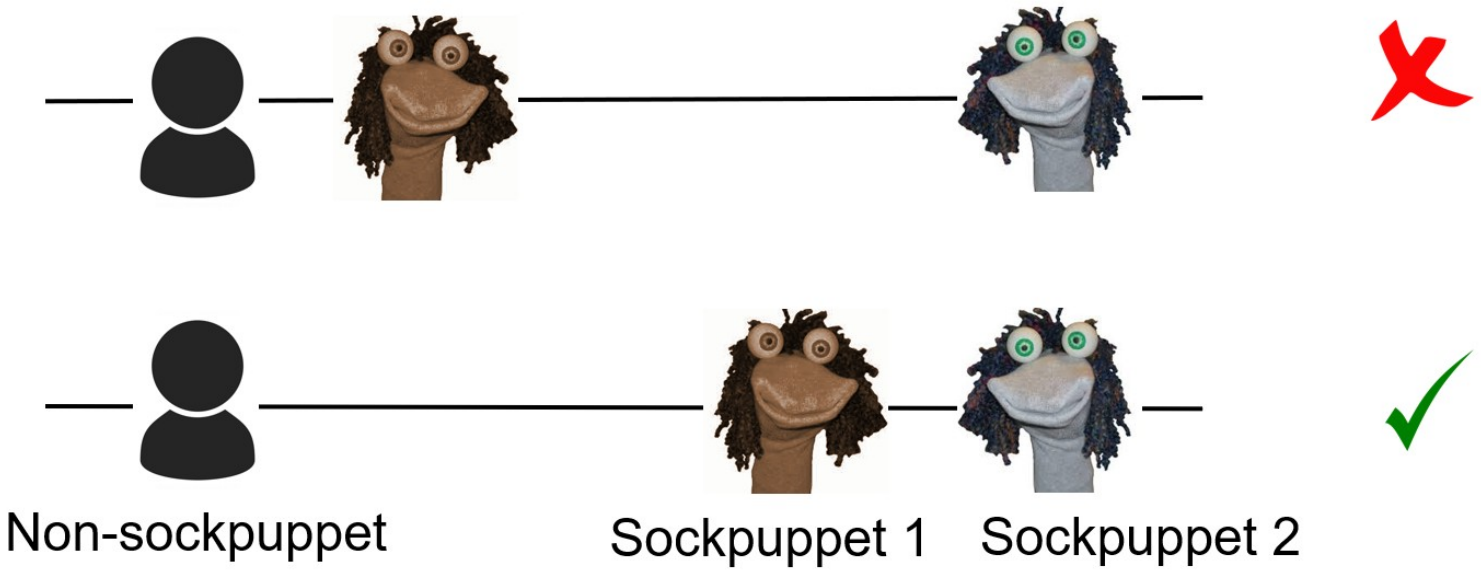
Alternate hypothesis:
Puppetmaster operates both sockpuppets similarly



Similarity is measured as cosine similarity between user posts' features: LIWC, sentiment, number of words, etc.

@vssubrah
vs@dartmouth.edu

Alternate Hypothesis wins



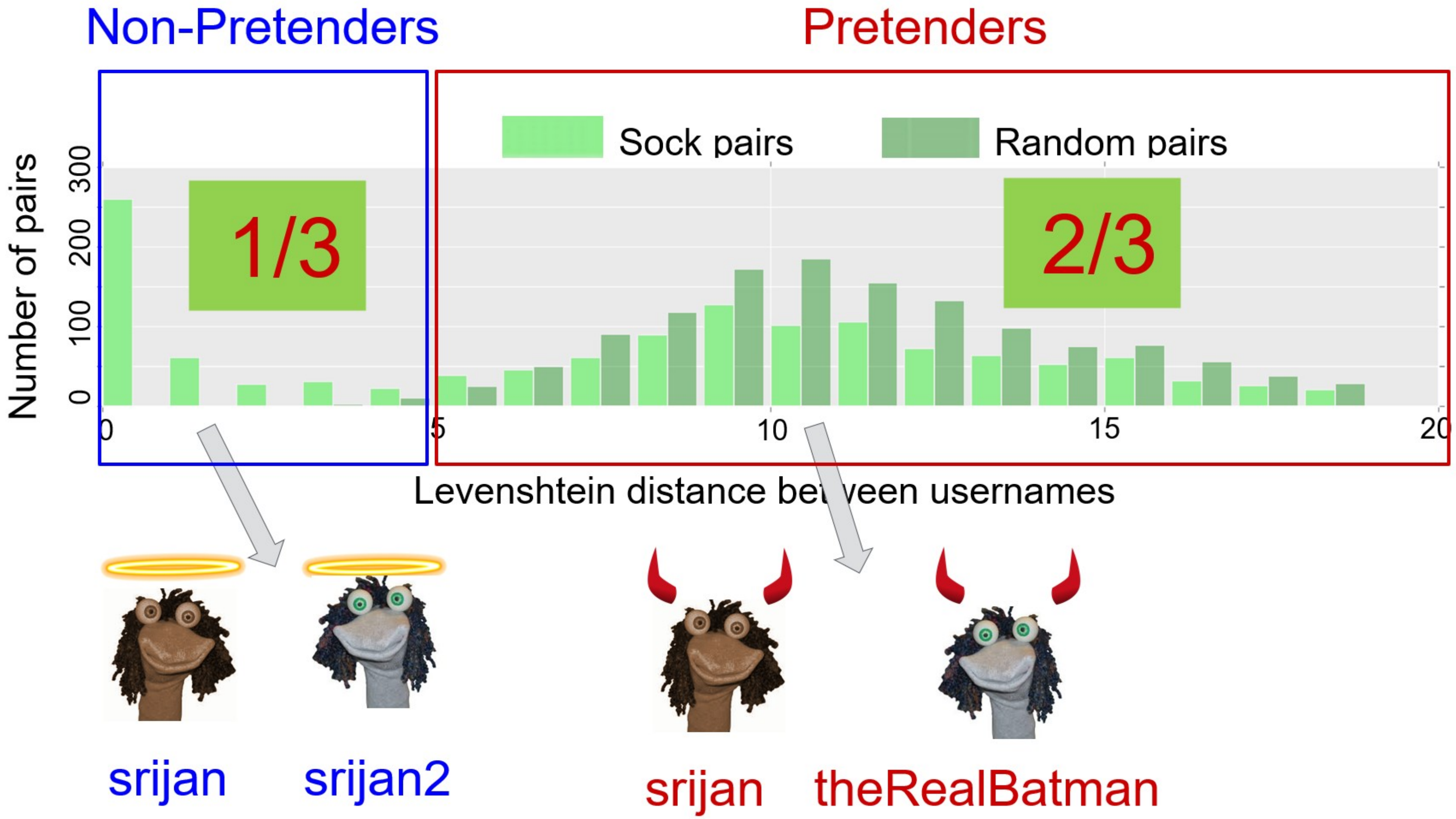
"Good sock/Bad sock" not common

Both sockpuppets are more similar to each other

$p < 10^{-3}$

@vssubrah
vs@dartmouth.edu

Are socks intended to be deceptive?



@vssubrah
vs@dartmouth.edu

Pretender vs. Non-Pretender Behavior

@vssubrah
vs@dartmouth.edu

Sockpuppet Types: Neutral

We quantify the amount of support by counting assenting, negation and dissenting words from LIWC

srijan Feb 5 2013, 3PM
best article ever!

theRealBatman Feb 5 2013, 3PM
why so?

60%
Neutral

@vssubrah
vs@dartmouth.edu

Sockpuppet Types: Supporting

We quantify the amount of support by counting assenting, negation and dissenting words from LIWC

srijan Feb 5 2013, 3PM
best article ever!

theRealBatman Feb 5 2013, 3PM
Totally agree!!

60% Neutral

30% Supporter

@vssubrah
vs@dartmouth.edu

Sockpuppet Types: Dissenting

We quantify the amount of support by counting assenting, negation and dissenting words from LIWC

srijan Feb 5 2013, 3PM
best article ever!

theRealBatman Feb 5 2013, 3PM
I don't think so

60%	30%	10%
Neutral	Supporter	Dissenter

@vssubrah
vs@dartmouth.edu

Predicting Socks: Features

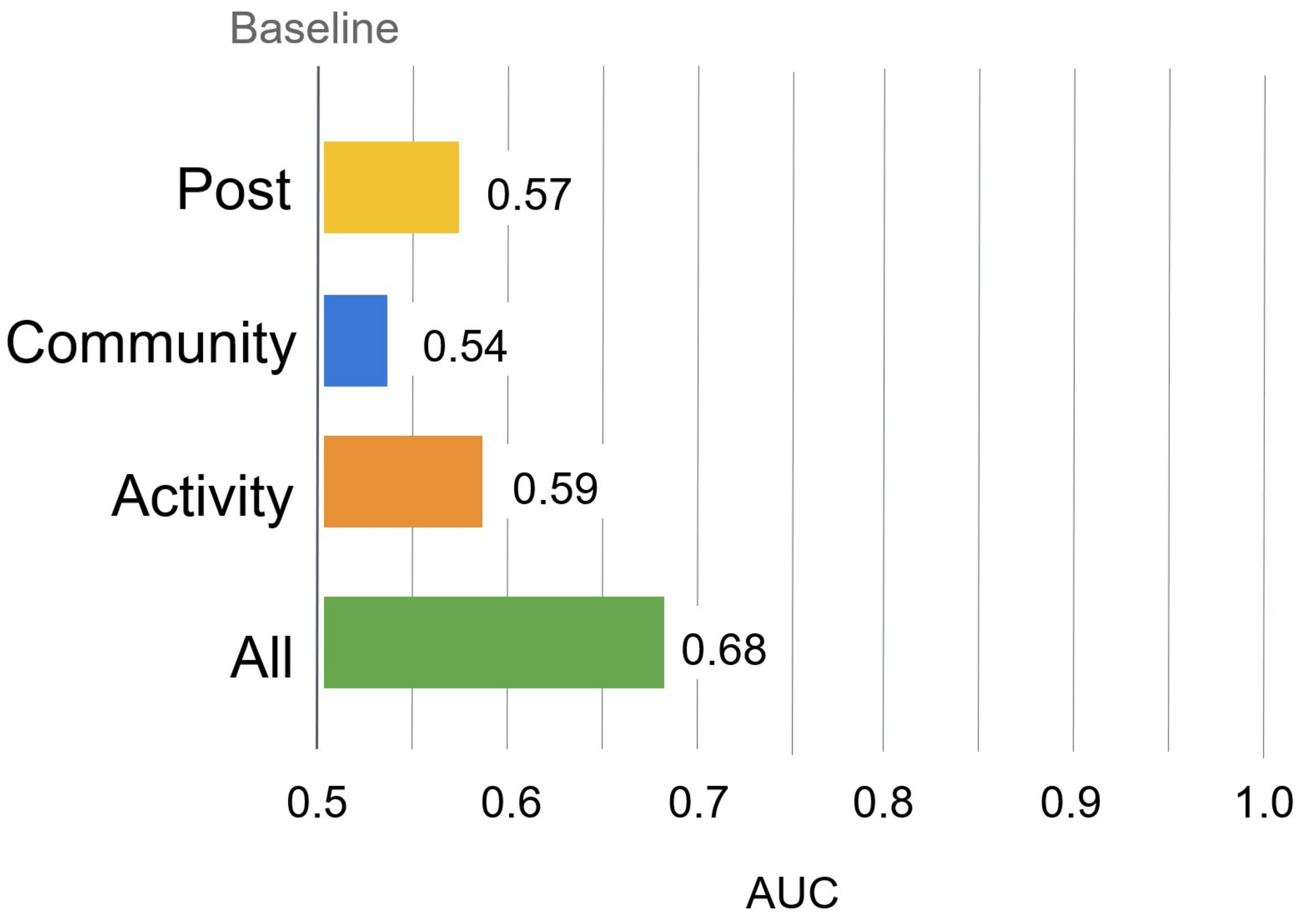
Activity-based
Number of posts
number of replies
reciprocity of posts
age of account
...

Post-based
Number of words
characters
LIWC counts
Readability
Sentiment
...

Community-based
Number of upvotes
Number of downvotes
...

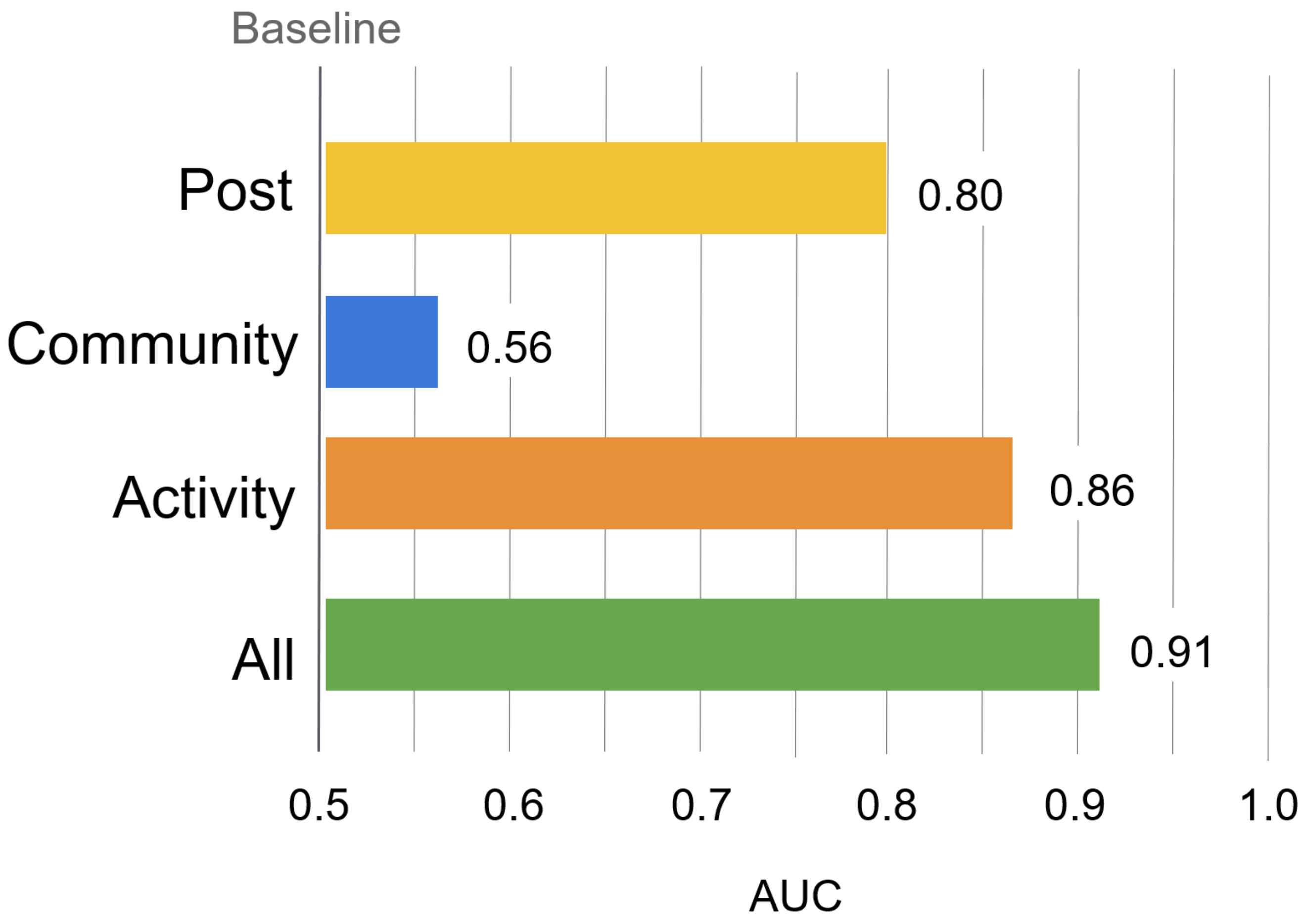
@vssubrah
vs@dartmouth.edu

Predicting Socks: Is Account A a Sock?



@vssubrah
vs@dartmouth.edu

Predicting Socks: Are accounts A,B a sock pair?



@vssubrah
vs@dartmouth.edu

Outline of talk

- Online Marketplaces: Review Fraud
- News & Other Discussion Forms: Sockpuppet Accounts
- **Wikipedia: Vandals**
- Twitter: Bots
- Malicious Actors – The Next Generation

VEWS: A Wikipedia Vandal Early Warning System.
S. Kumar, F. Spezzano and V.S. Subrahmanian. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2015.

@vssubrah
vs@dartmouth.edu

Vandals at Work

~ 7% edits involve vandalism

~ 3-4 % editors are vandals

@vssubrah
vs@dartmouth.edu

VEWS Data

34,000 Editors Half are vandals

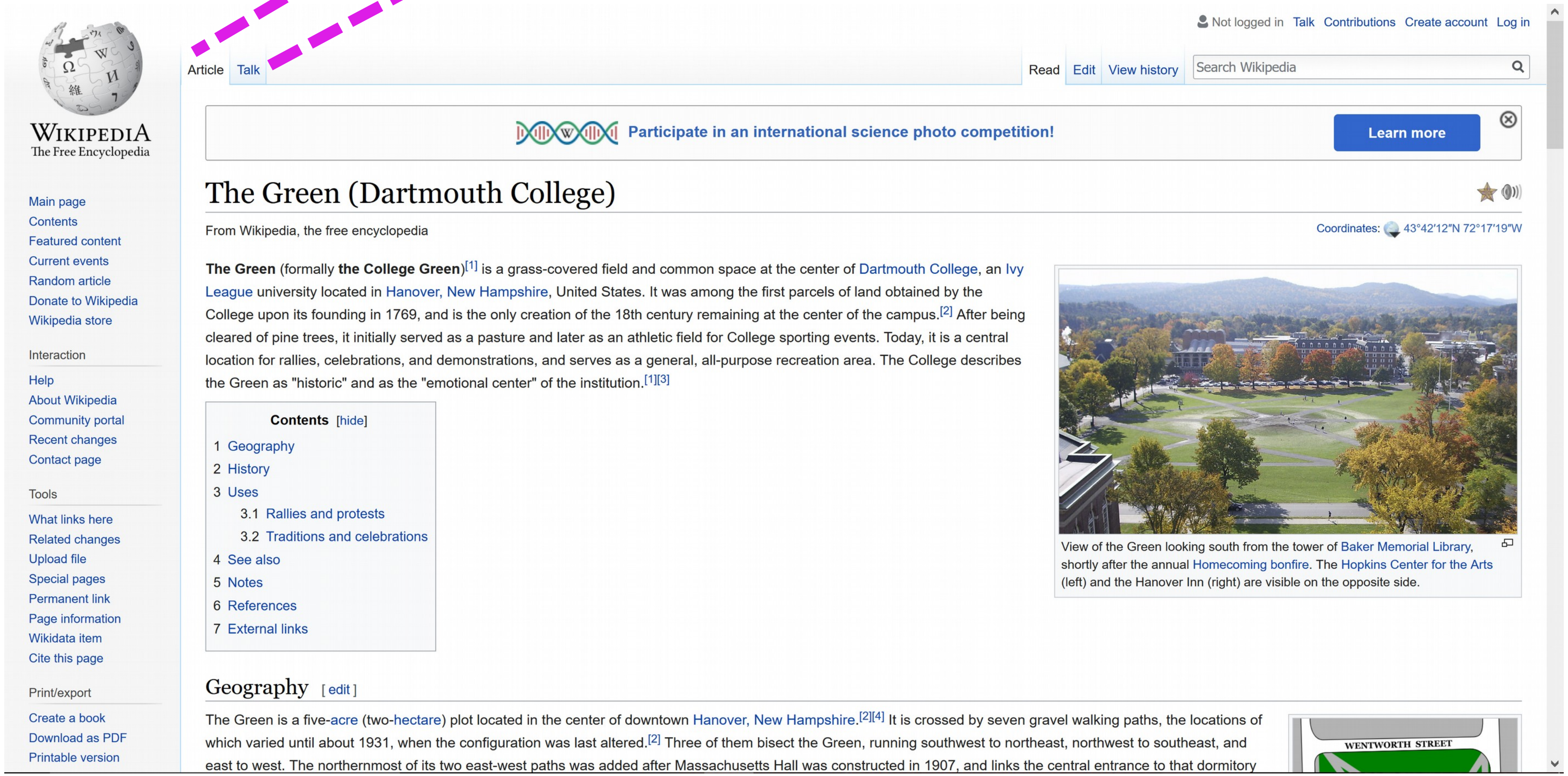
770,000 Edits 160,000 edits by vandals

Time: Jan 2013 - July 2014

Data available at: https://www.cs.umd.edu/~vs/vews/

@vssubrah
vs@dartmouth.edu

Wikipedia Pages: Article & Talk



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Create a book
Download as PDF
Printable version

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Participate in an international science photo competition! Learn more

The Green (Dartmouth College)

From Wikipedia, the free encyclopedia

Coordinates: 43°42′12″N 72°17′19″W

The Green (formally **the College Green**)[1] is a grass-covered field and common space at the center of Dartmouth College, an Ivy League university located in Hanover, New Hampshire, United States. It was among the first parcels of land obtained by the College upon its founding in 1769, and is the only creation of the 18th century remaining at the center of the campus.[2] After being cleared of pine trees, it initially served as a pasture and later as an athletic field for College sporting events. Today, it is a central location for rallies, celebrations, and demonstrations, and serves as a general, all-purpose recreation area. The College describes the Green as "historic" and as the "emotional center" of the institution.[1][3]

View of the Green looking south from the tower of Baker Memorial Library, shortly after the annual Homecoming bonfire. The Hopkins Center for the Arts (left) and the Hanover Inn (right) are visible on the opposite side.

Contents [hide]

1 Geography
2 History
3 Uses
3.1 Rallies and protests
3.2 Traditions and celebrations
4 See also
5 Notes
6 References
7 External links

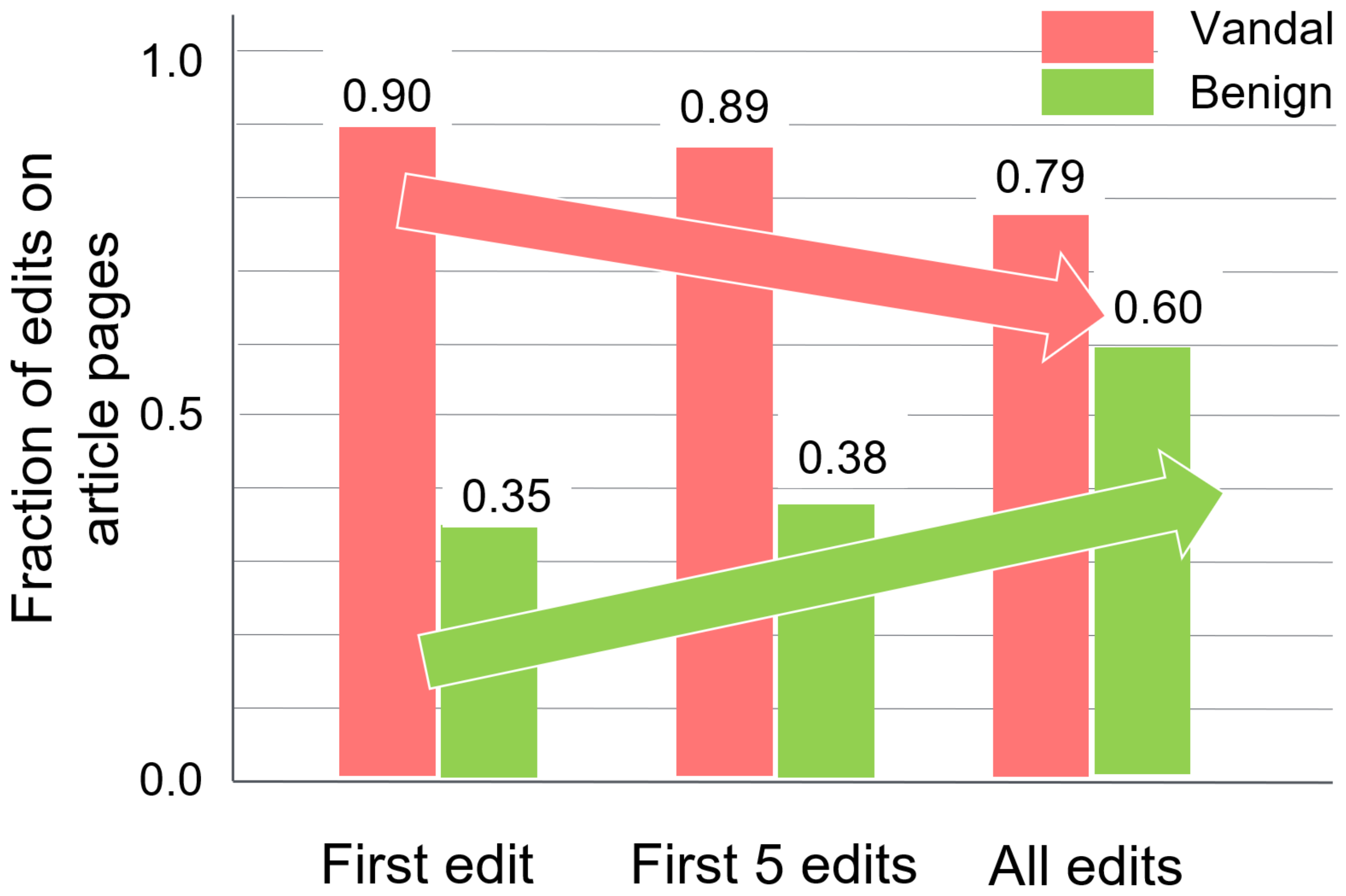
Geography [edit]

The Green is a five-acre (two-hectare) plot located in the center of downtown Hanover, New Hampshire.[2][4] It is crossed by seven gravel walking paths, the locations of which varied until about 1931, when the configuration was last altered.[2] Three of them bisect the Green, running southwest to northeast, northwest to southeast, and east to west. The northernmost of its two east-west paths was added after Massachusetts Hall was constructed in 1907, and links the central entrance to that dormitory

WENTWORTH STREET

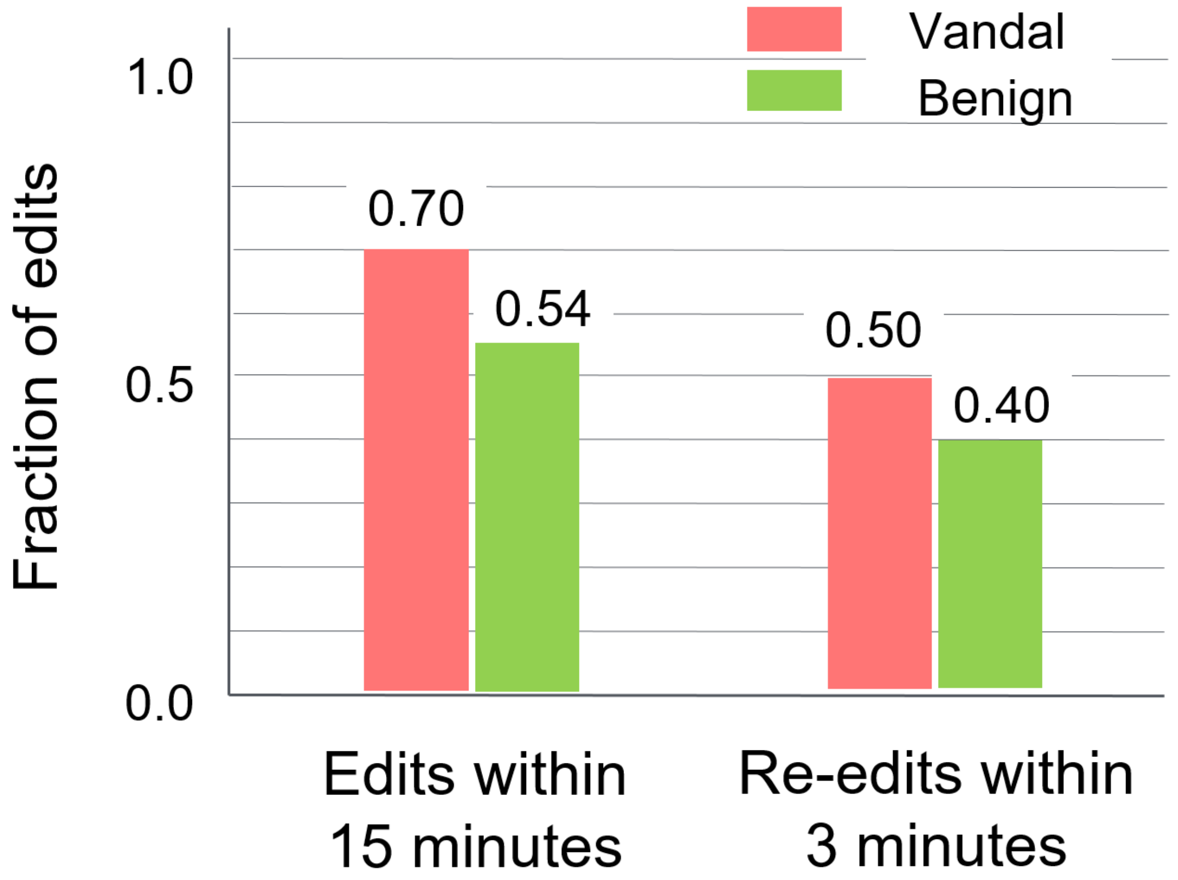
@vssubrah
vs@dartmouth.edu

Vandals rarely talk to others!



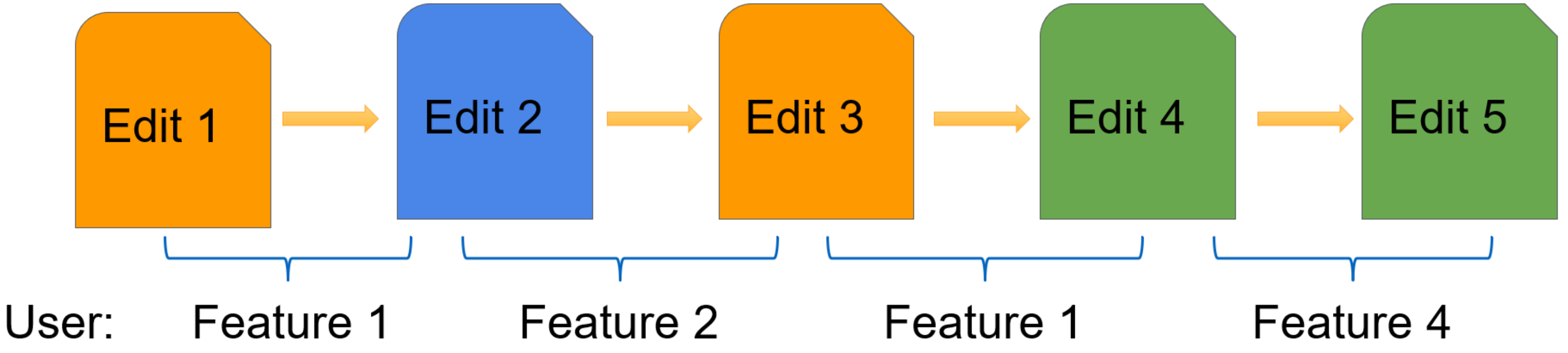
@vssubrah
vs@dartmouth.edu

Vandals edit in rapid-fire mode



@vssubrah
vs@dartmouth.edu

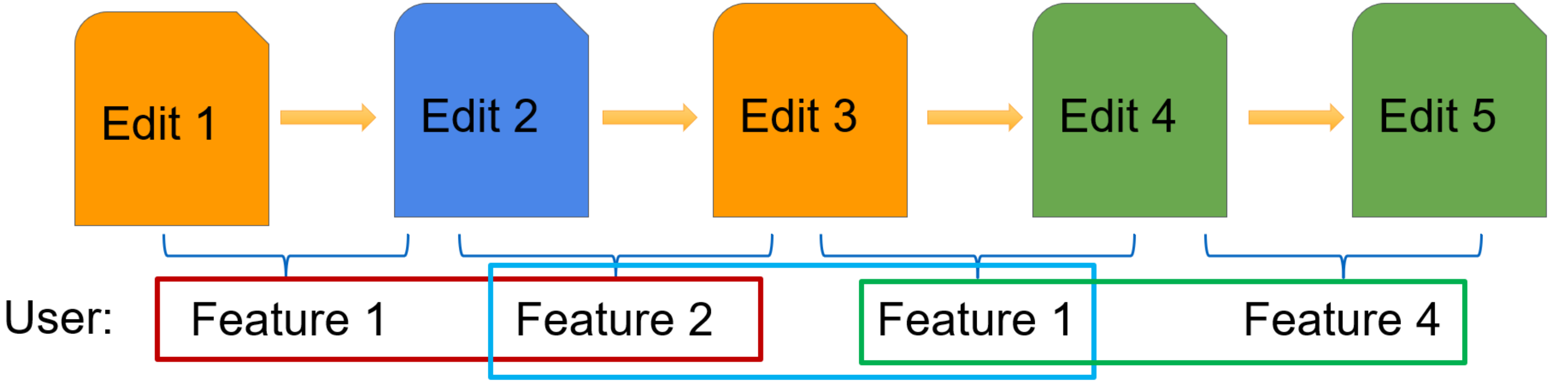
Pairwise edit features



Time x Type of page x First edit x Distance x Similarity
x Reverted or not

@vssubrah
vs@dartmouth.edu

Transition Features



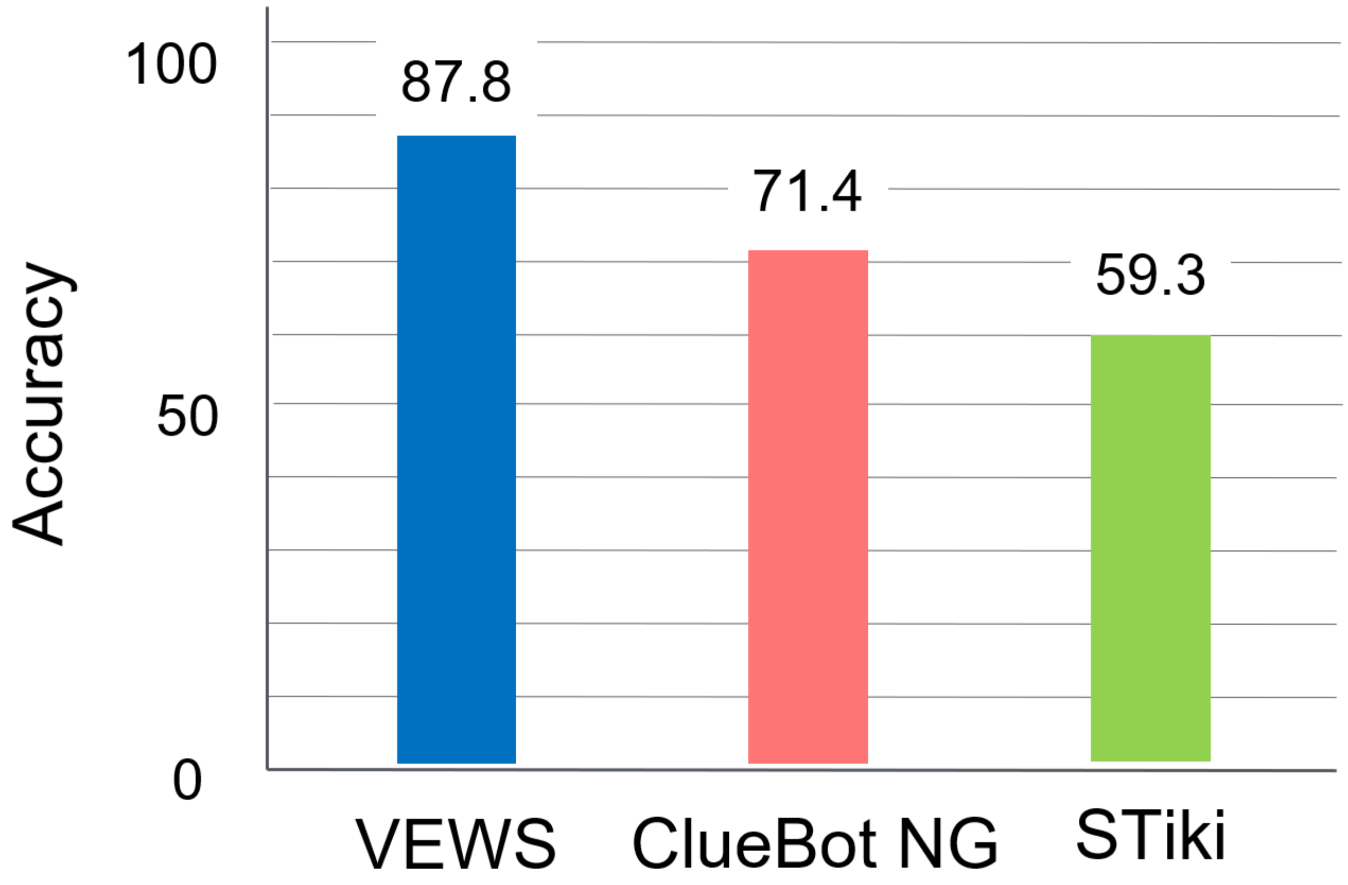
	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Feature 1		0.5		0.5	
Feature 2	1				
Feature 3					
Feature 4					
Feature 5					

N x N

X[i,j] = probability that feature vector j occurs immediately after feature vector i

@vssubrah
vs@dartmouth.edu

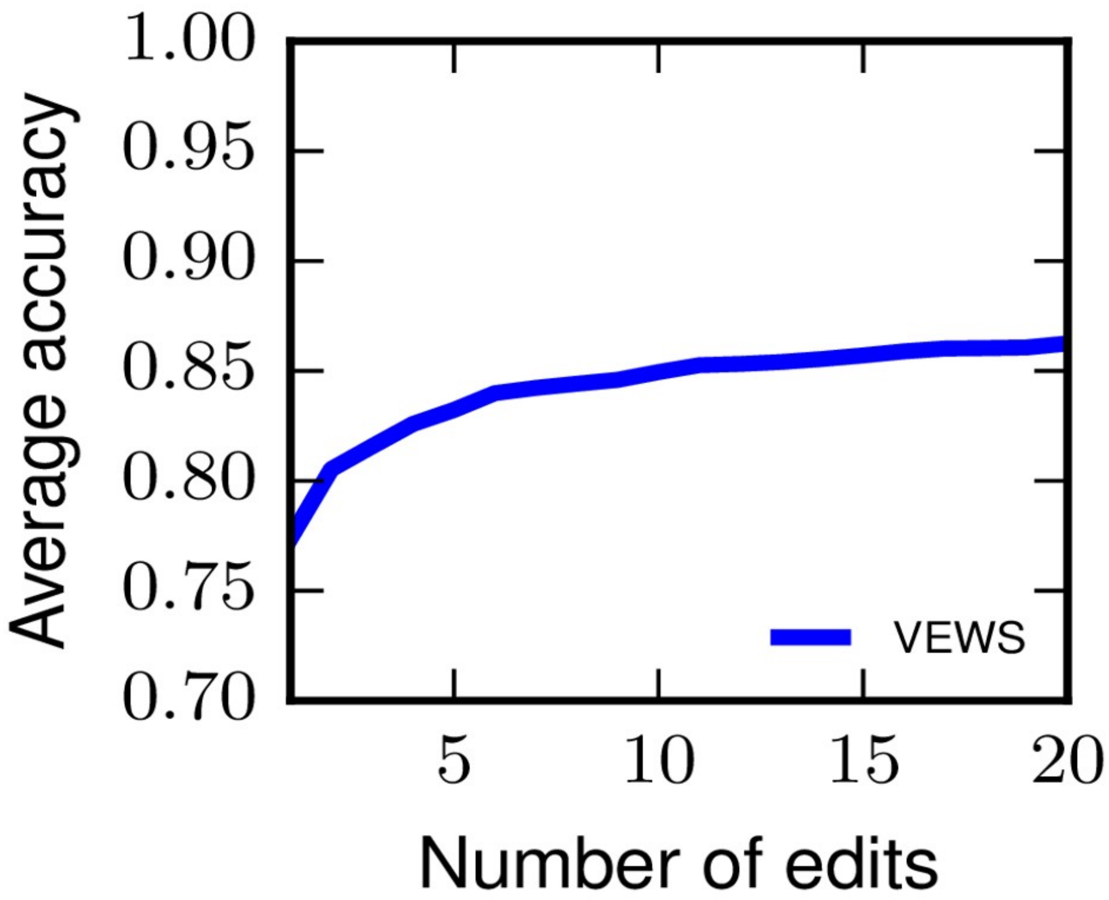
VEWS Predictive Accuracy



- VEWS identifies 87% of vandals on or before first reversion.
- 44% of vandals are identified before first reversion.

@vssubrah
vs@dartmouth.edu

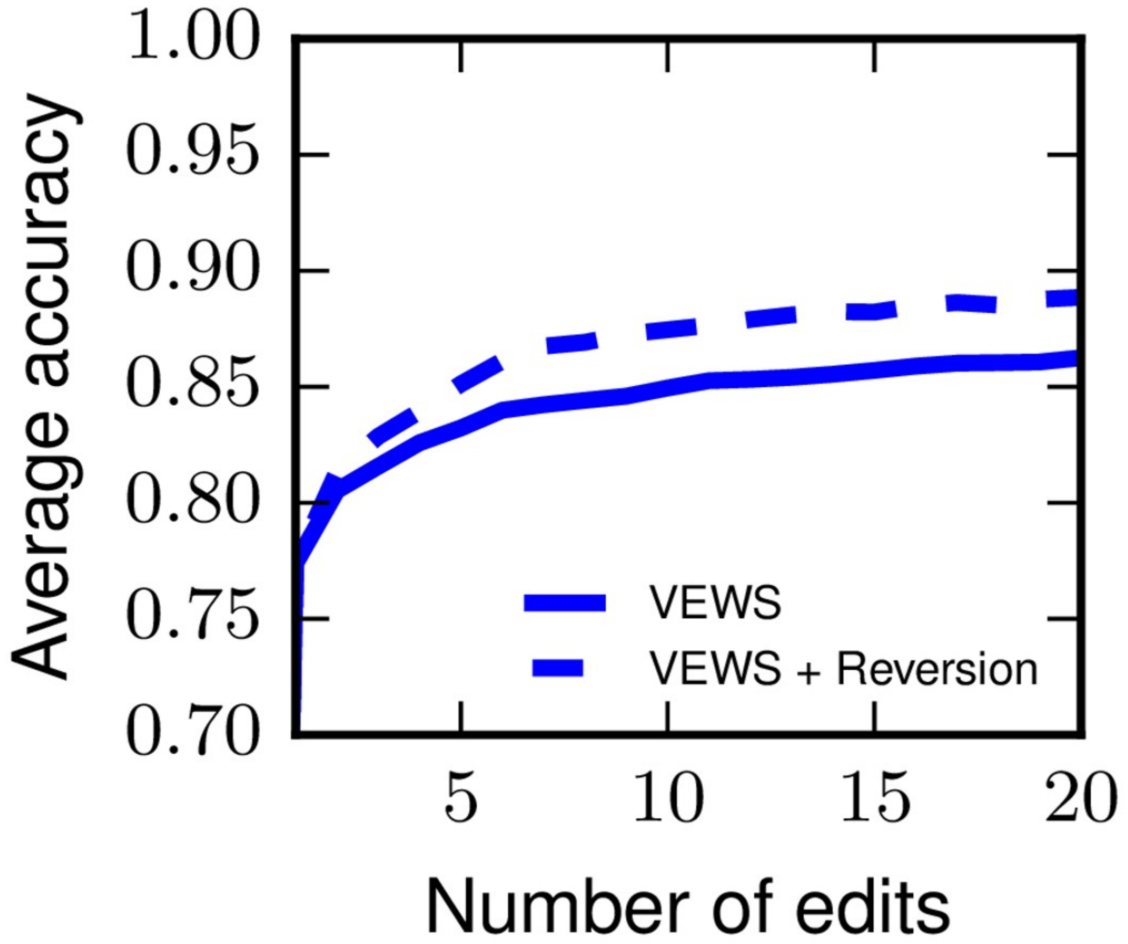
VEWS' Speed in Identifying Vandals



VEWS identifies vandals (on average) in 2.13 edits.

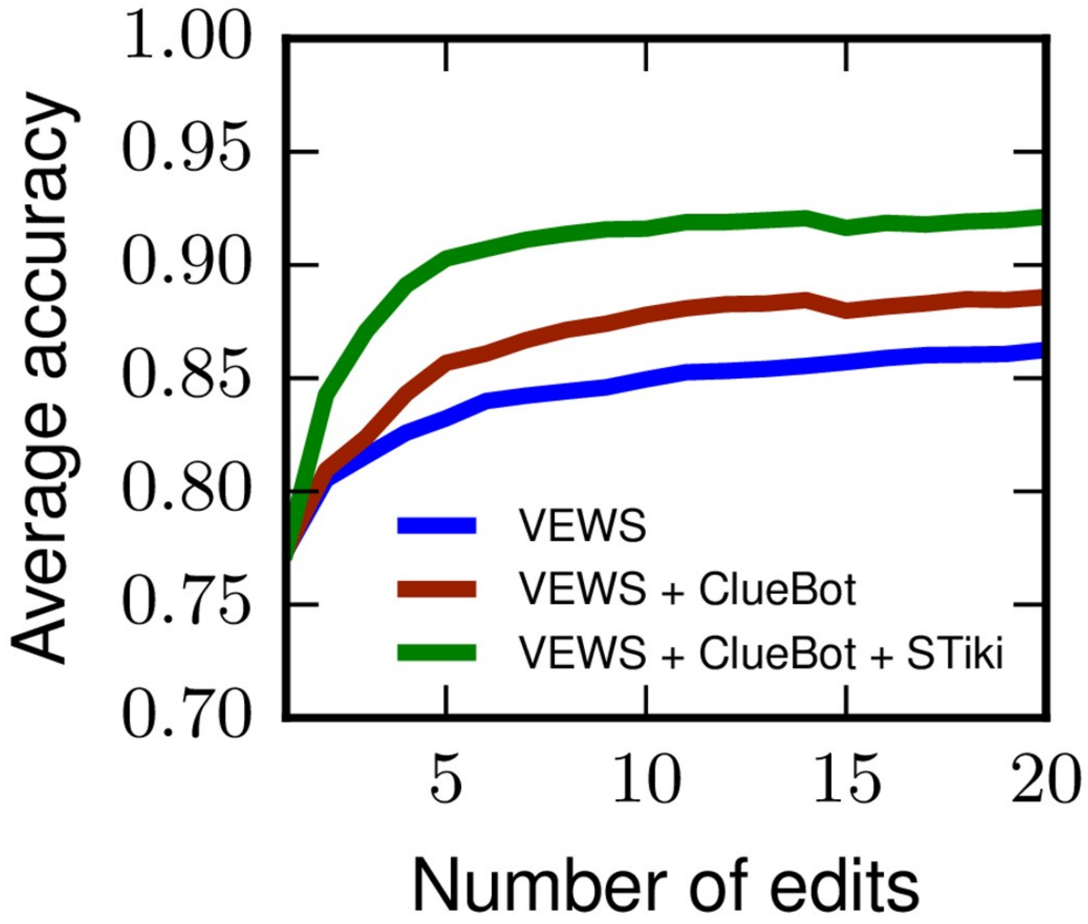
@vssubrah
vs@dartmouth.edu

Reversion Information Helps (a little)



@vssubrah
vs@dartmouth.edu

Combining with Past Work Helps



@vssubrah
vs@dartmouth.edu

Outline of talk

- Online Marketplaces: Review Fraud
- News & Other Discussion Forms: Sockpuppet Accounts
- Wikipedia: Vandals
- **Twitter: Bots**
- Malicious Actors – The Next Generation

V.S. Subrahmanian et al. "The DARPA Twitter bot challenge." *Computer* 49.6 (2016): 38-46.

Dickerson, John P., Vadim Kagan, and V. S. Subrahmanian. "Using sentiment to detect bots on twitter: Are humans more opinionated than bots?." *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 2014.

@vssubrah
vs@dartmouth.edu

Bots in the 2014 India Election

- Largest democratic election in human history
- Tracked 31 topics (national politicians, political parties) over 10 month period
- Over
  - 17M users
  - 25M posts (after eliminating irrelevant posts from a ~600M tweet data set)
  - 45M edges

@vssubrah
vs@dartmouth.edu

Features

Tweet Syntax
- #hashtags, #mentions, #links, etc

Tweet Semantics
- Sentiment related features for user

User Behavior
- Tweet spread/frequency/repeats/geo
- Tweet volume histograms by topic
- Sentiment: normalized flip flops(t), variance(t), monthly variance(t)

User Neighborhood (and behavior)
- Multiple measures looking at agreement/ disagreement between user sentiments and those of people in his neighborhood

Contradiction Rank
- $CR(u,t) = x_t^+ y_t^- + x_t^- y_t^+$ where
  - x_t^+ is the fraction of u's tweets with sentiment that are positive w.r.t. t
  - y_t^+ is the fraction of all tweets [not just u's] with sentiment that are positive w.r.t. t
  - x_t^-, y_t^- defined similarly

Agreement Rank $AR(u,t) = x_t^+ y_t^+ + x_t^- y_t^-$

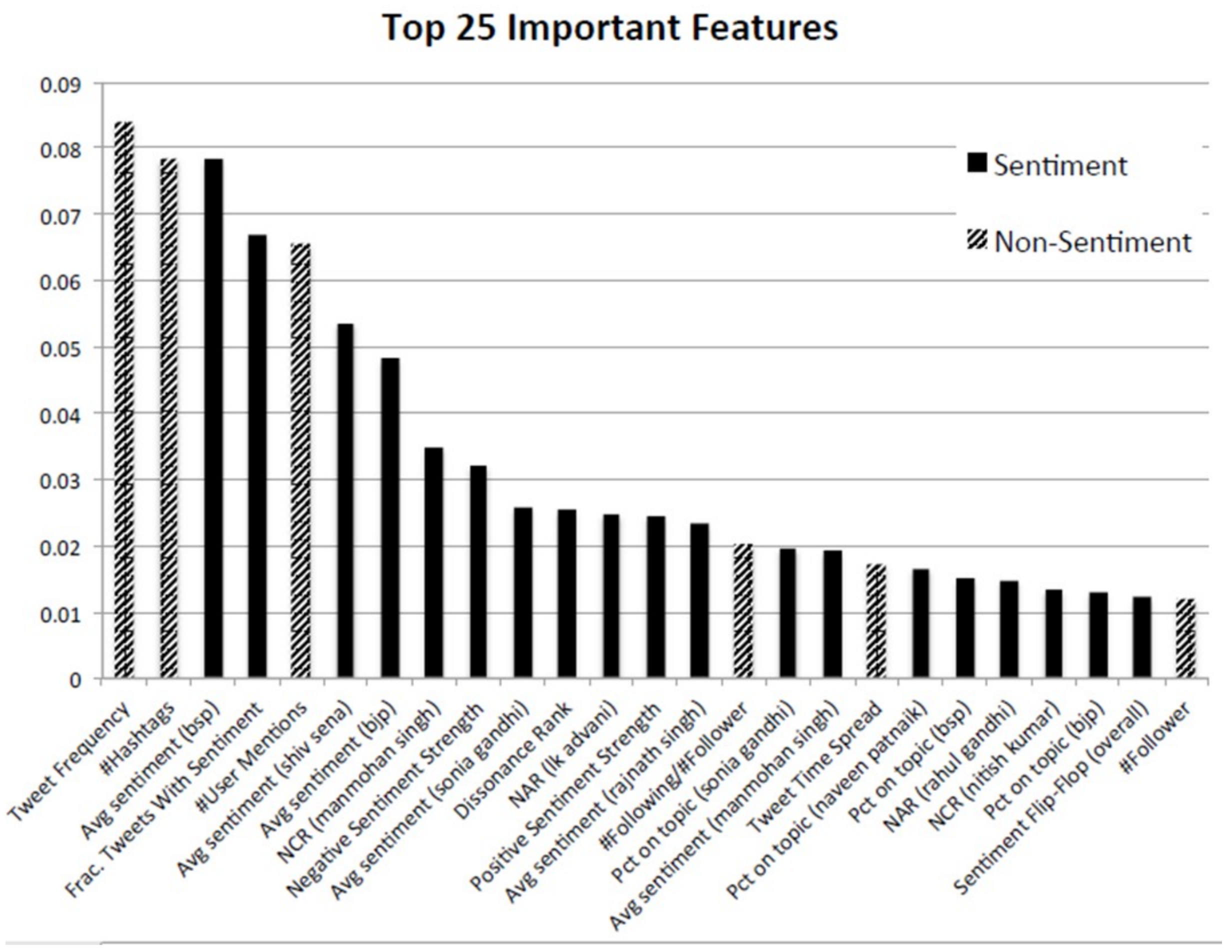
Dissonance rank $DR(u) = \sum_{t \in TOI} \frac{CR(u,t)}{AR(u,t)}$

Positive Sentiment Strength
- Average sentiment score (for t) from u's tweets that are positive about t

+/- Sentiment Polarity Fraction
- Percentage of u's tweets on t that are positive/negative

@vssubrah
vs@dartmouth.edu

Bots vs. Humans

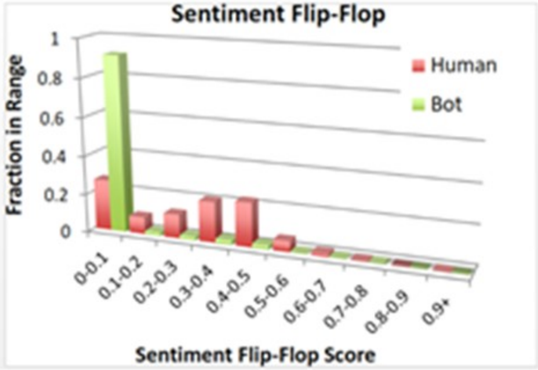


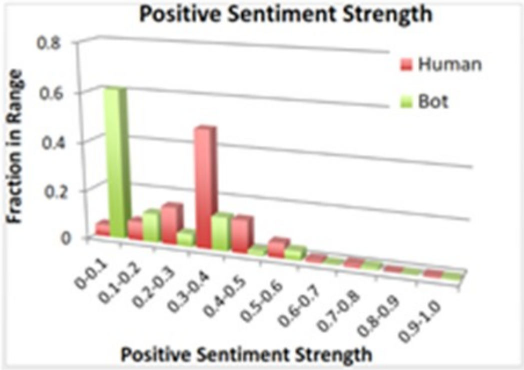
Bots vs. Humans?

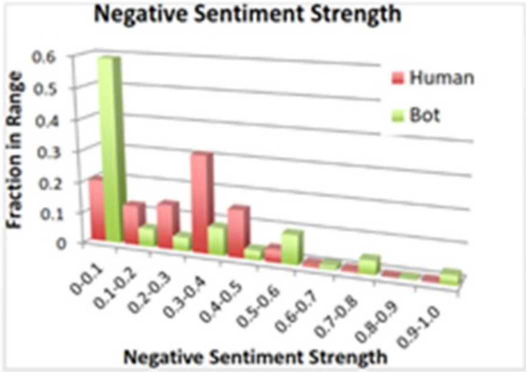
- Who flip flops more?
- Whose positive opinions are stronger?
- Whose negative opinions are stronger?
- Who tend to write more tweets with sentiment?
- Who tend to disagree more?

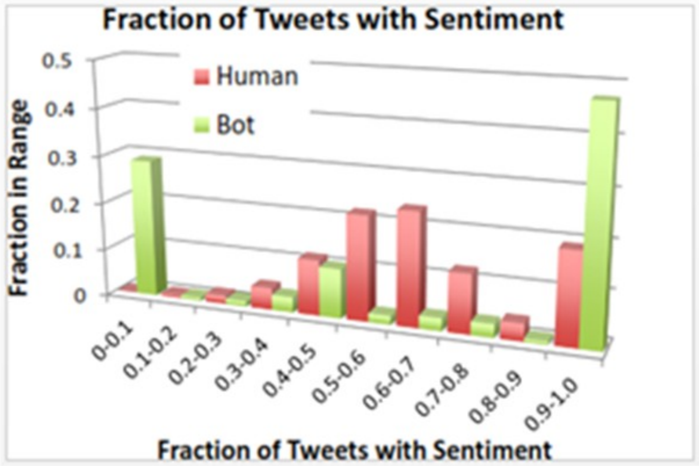
@vssubrah
vs@dartmouth.edu

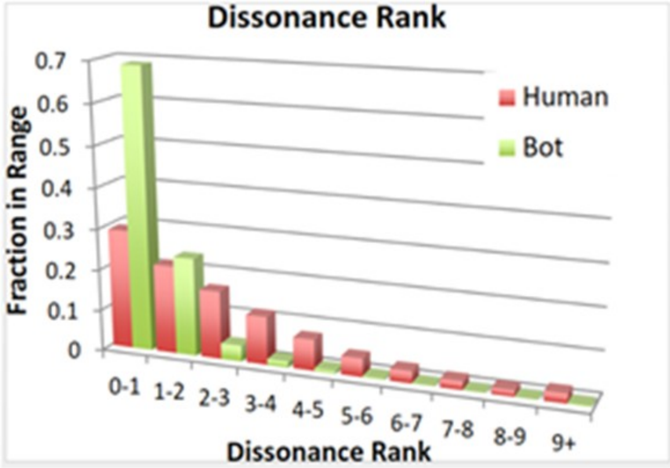
Bots vs. Humans











@vssubrah
vs@dartmouth.edu

DARPA Twitter Bot Chall

Phase 1 (9 days)
Identify an initial set of bots through assessment of how bot developers would operate.
66 features

Phase 2 (3 days)
Semi-supervised clustering with different similarity functions & outlier detection
Total about ~125 features

Phase 3 (4 days)
Straight machine learning using SVM and Random Forest ensemble
Total about ~175 features

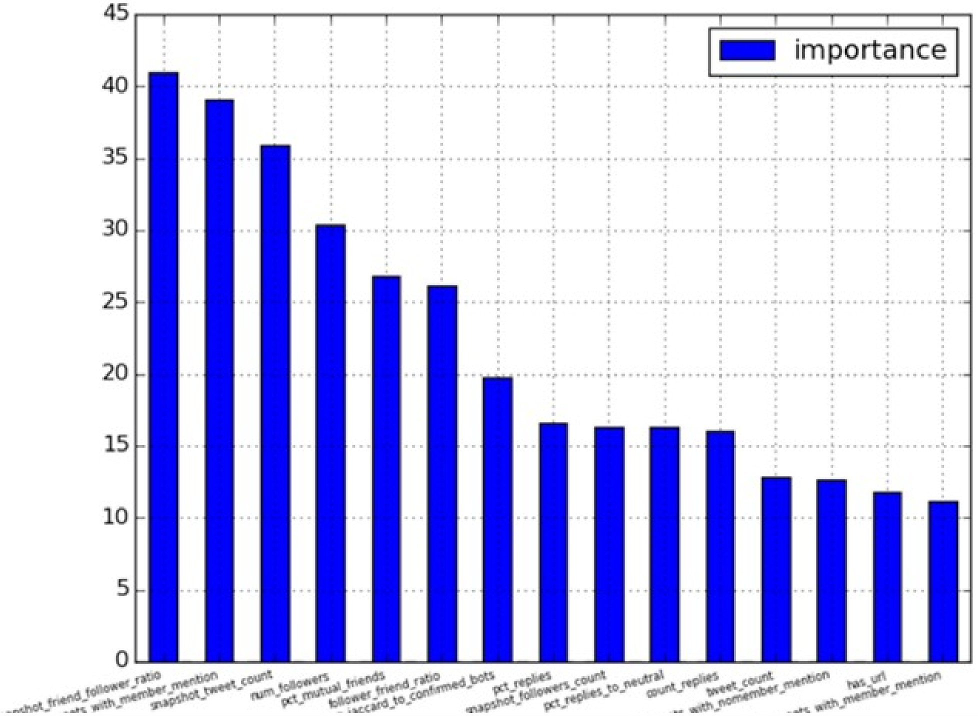


TABLE 1. Results of the DARPA Twitter Bot Challenge.

Team	Misses	Hits	Guesses	Accuracy	Speed	Final score
SentiMetrix	1	39	40	38.75	12	50.75
University of Southern California	0	39	39	39.00	6	45.00
DESPIC	7	39	46	37.25	6	43.25
IBM	4	39	43	38.00	5	43.00
Boston Fusion	9	39	48	36.75	5	41.75
Georgia Tech	56	38	94	24.00	0	24.00

The accuracy column is the value ($h - 0.25m$), where h is the number of hits (correct guesses) and m is the number of misses (incorrect guesses). The speed column equals the number of days remaining in the challenge after the team had discovered all bots. DESPIC is the Indiana University/University of Michigan team. For each team t, FinalScore(t) = Hits(t) – 0.25 × Misses(t) + Speed.

@vssubrah
vs@dartmouth.edu

Outline of talk

- Online Marketplaces: Review Fraud
- News & Other Discussion Forms: Sockpuppet Accounts
- Wikipedia: Vandals
- Twitter: Bots
- **Malicious Actors – The Next Generation**

@vssubrah
vs@dartmouth.edu

Malicious Actors on Social Platforms: The Future

- **Cross platform** Coordinated attacks across multiple platforms
- **Distributed, low key** Low key activities within each platform
- **Conformity** Greater conformance with opinion within local communities with small efforts to shift opinion
- **Greater engagement** of bots and malicious actors with existing communities online
- **Combination with traditional cyber methods** Combine social attacks with more traditional hacks

Contact Information

V.S. Subrahmanian
Dept. of Computer Science
Dartmouth College
Hanover, NH 03755
vs@dartmouth.edu
@vssubrah
www.cs.dartmouth.edu/vs/