A Weakly Supervised Approach for Adaptive Detection of Cyberbullying Roles

Bert Huang
Department of Computer Science
Virginia Tech

CyberSafety Workshop 10/28/16

Cyberbullying

Cyberbullying

- Cyberbullying: “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices”

Cyberbullying

- Cyberbullying: “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices”
- Forms of cyberbullying:

Cyberbullying

- Cyberbullying: “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices”
- Forms of cyberbullying:
  - Offensive and negative comments, name calling, rumor spreading, threats, public shaming

Cyberbullying

- Cyberbullying: “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices”
- Forms of cyberbullying:
  - Offensive and negative comments, name calling, rumor spreading, threats, public shaming
- Linked to mental health issues, e.g., depression, suicide

Cyberbullying

- Cyberbullying: “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices”
- Forms of cyberbullying:
  - Offensive and negative comments, name calling, rumor spreading, threats, public shaming
- Linked to mental health issues, e.g., depression, suicide
- Anytime, persistent, public, anonymous

Get Help Now

When you, your child, or someone close to you is being bullied, there are many steps to take to help resolve the situation. Make sure you understand what bullying is and what it is not, the warning signs of bullying, and steps to take for preventing and responding to bullying, including how to talk to children about bullying, prevention in schools and communities, and how to support children involved.

After reviewing that information, if you feel you have done everything you can to resolve the situation and nothing has worked, or someone is in immediate danger, there are ways to get help.

The problem	What you can do
There has been a crime or someone is at immediate risk of harm.	Call 911.
Someone is feeling hopeless, helpless, thinking of suicide.	Contact the National Suicide Prevention Lifeline online or at 1-800-273-TALK (8255). The toll-free call goes to the nearest crisis center in our national network. These centers provide 24-hour crisis counseling and mental health

Talk Plan

1. Challenges in Machine Learning for Cyberbullying
2. New Method for Weakly Supervised Learning for Detection
3. Open Problem: Automated Interventions

Collaborators

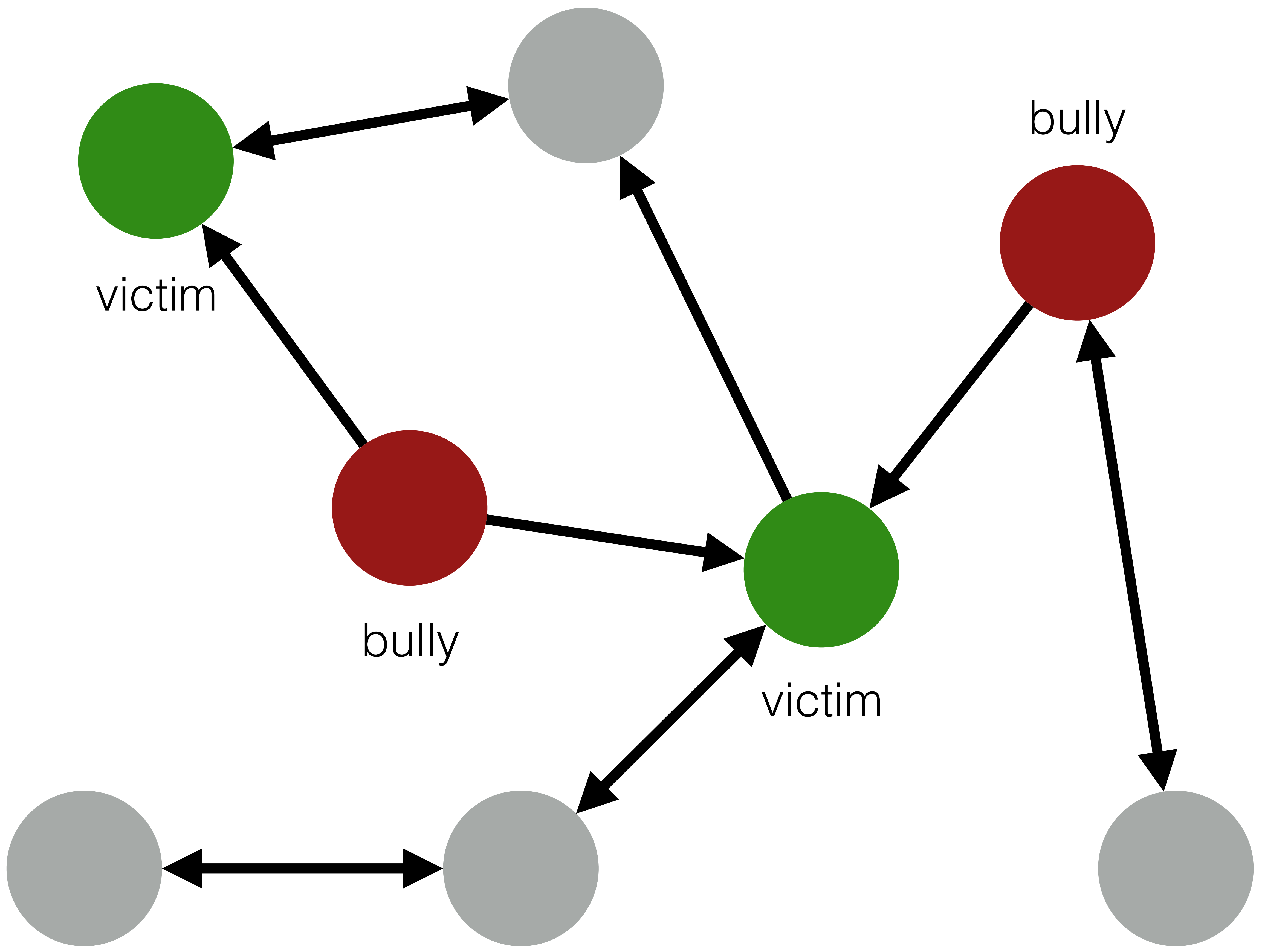
Elaheh Raisi
Ph.D. student
Dept. of Computer Science

James Hawdon
Director of the Center for Peace Studies and Violence Prevention
Dept. of Sociology

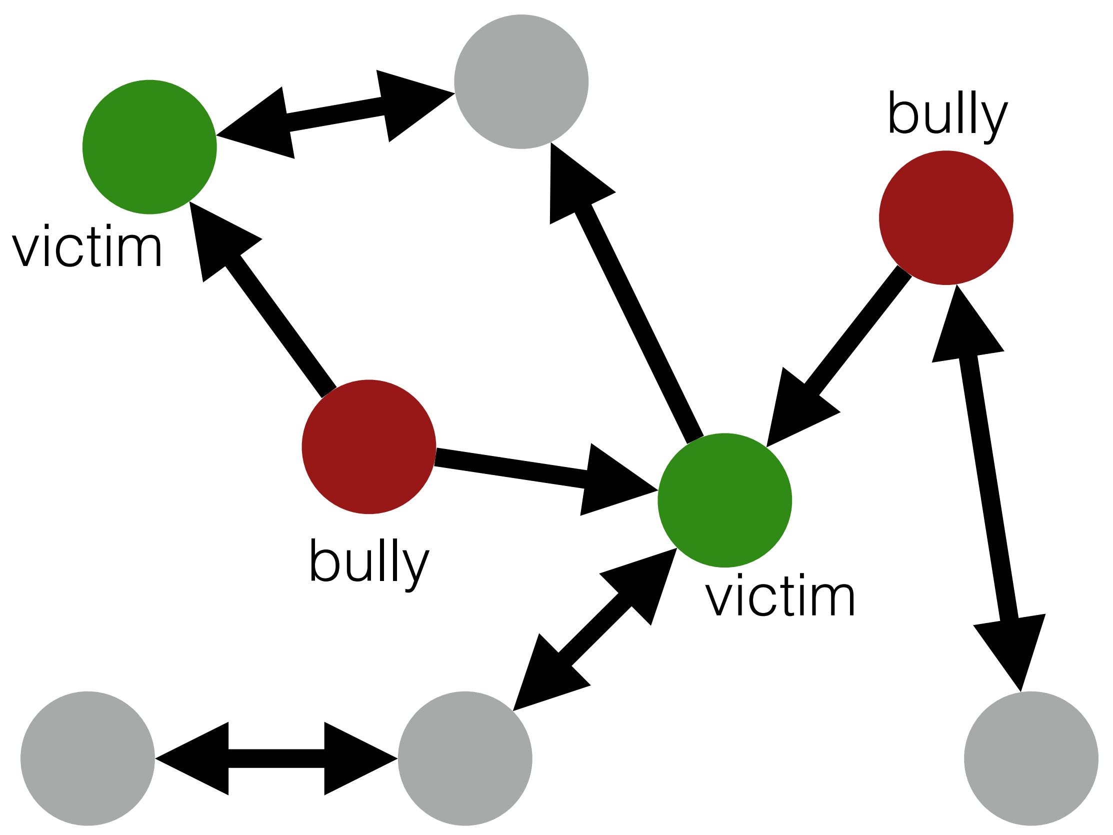
Anthony Peguero
Associate Professor
Dept. of Sociology

-1-

Challenges for Machine Learning of Cyberbullying Detectors

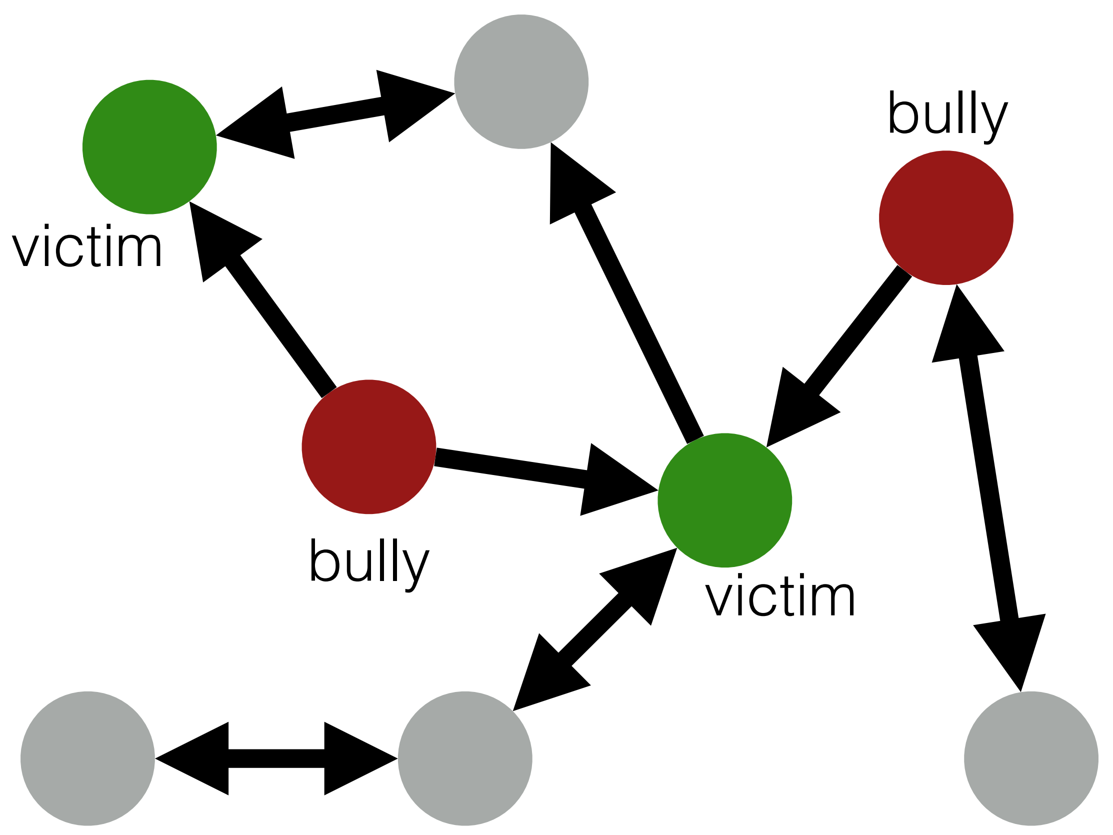

bully
victim
bully
victim

Challenges for Detecting Cyberbullying with Machine Learning



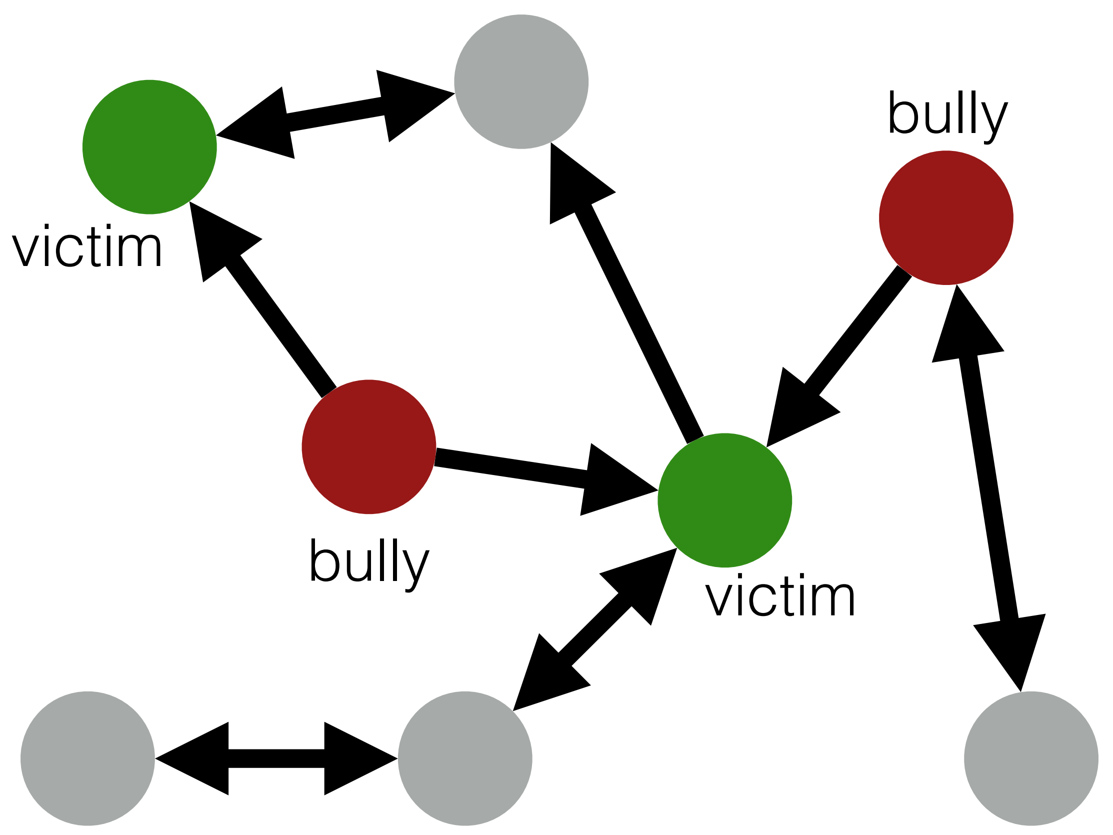
Challenges for Detecting Cyberbullying with Machine Learning

- Social structure is important



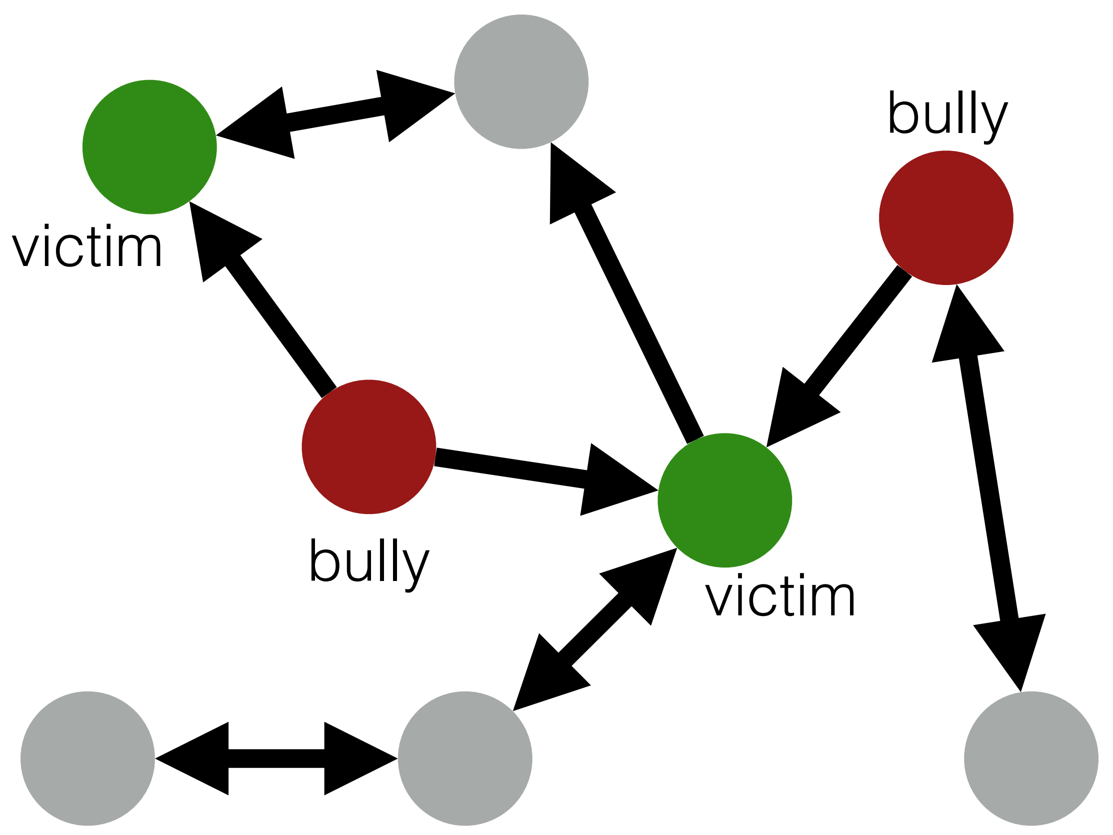
Challenges for Detecting Cyberbullying with Machine Learning

- Social structure is important
- Need scalable algorithms for massive data



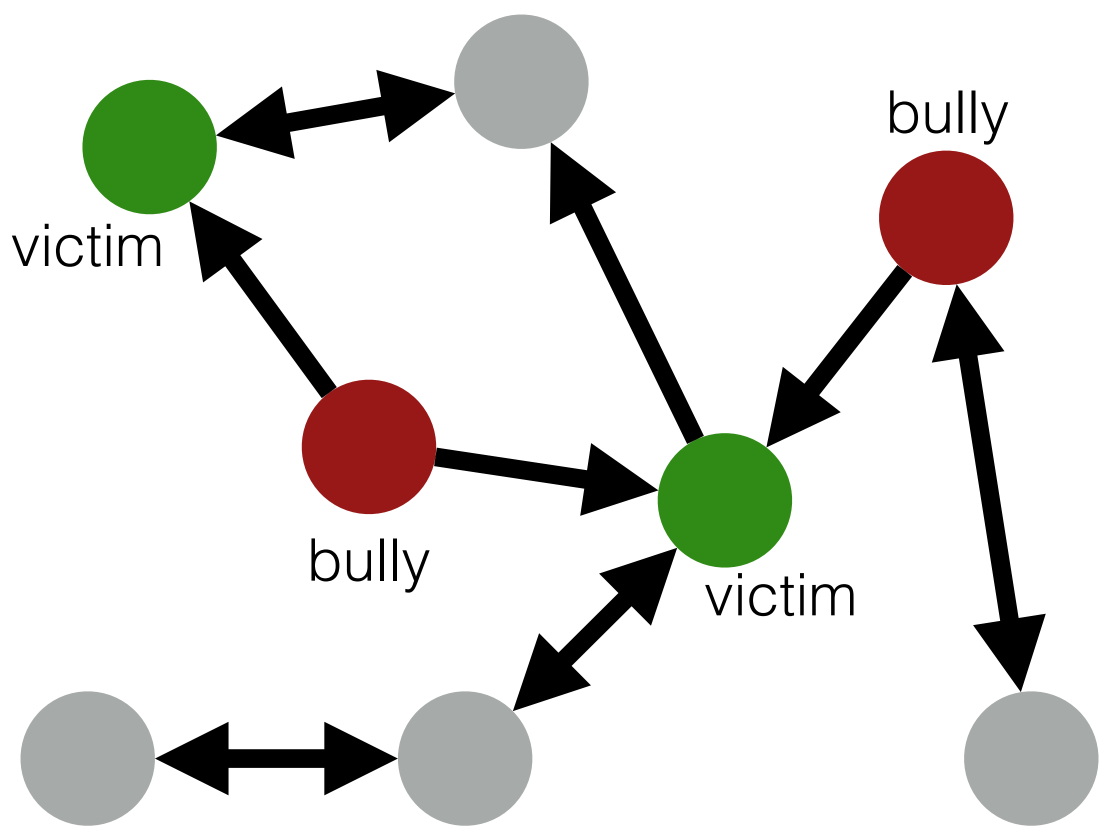
Challenges for Detecting Cyberbullying with Machine Learning

- Social structure is important
- Need scalable algorithms for massive data
- Language is changing:



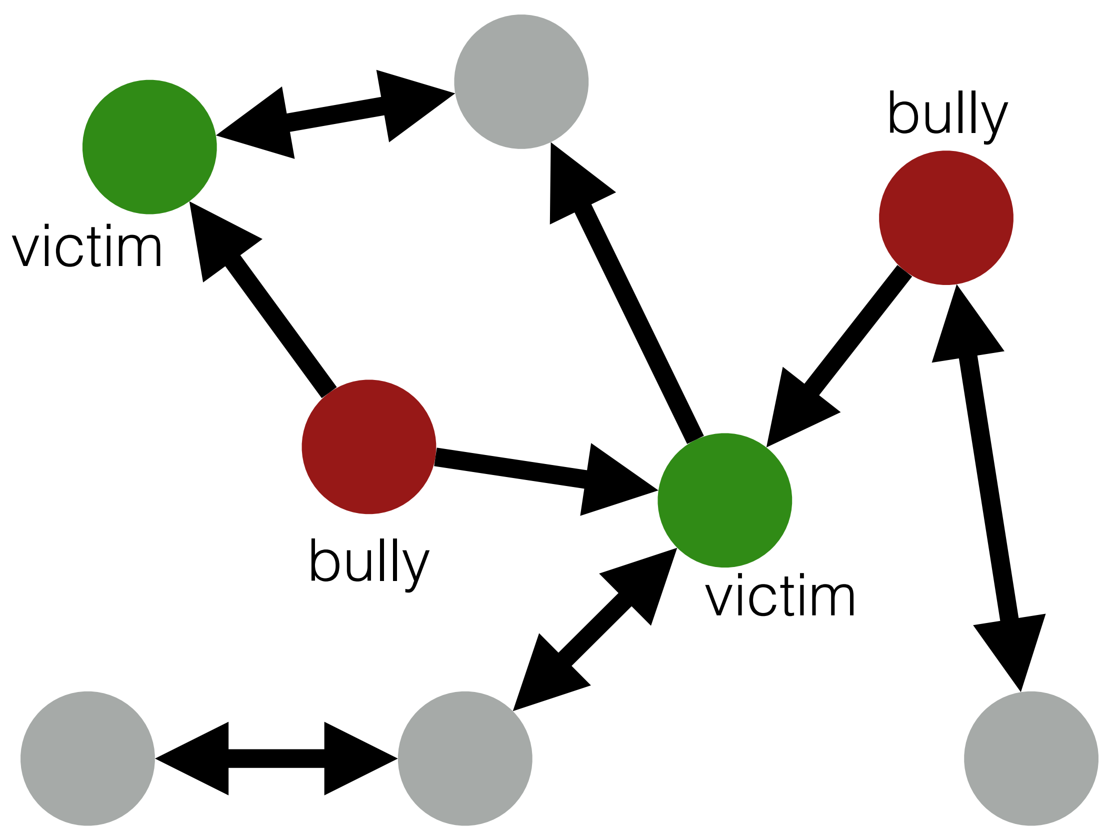
Challenges for Detecting Cyberbullying with Machine Learning

- Social structure is important
- Need scalable algorithms for massive data
- Language is changing:
  - New slang is frequently introduced or old slang becomes outdated



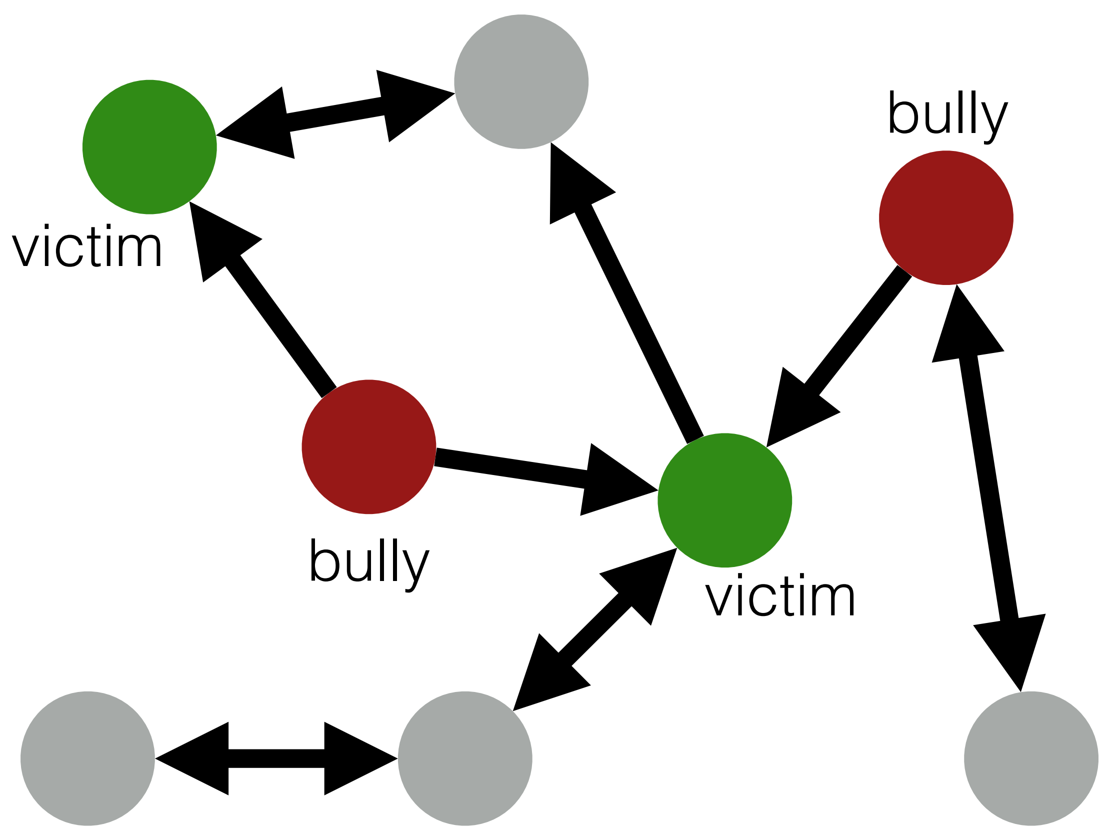
Challenges for Detecting Cyberbullying with Machine Learning

- Social structure is important
- Need scalable algorithms for massive data
- Language is changing:
  - New slang is frequently introduced or old slang becomes outdated
- Annotation:



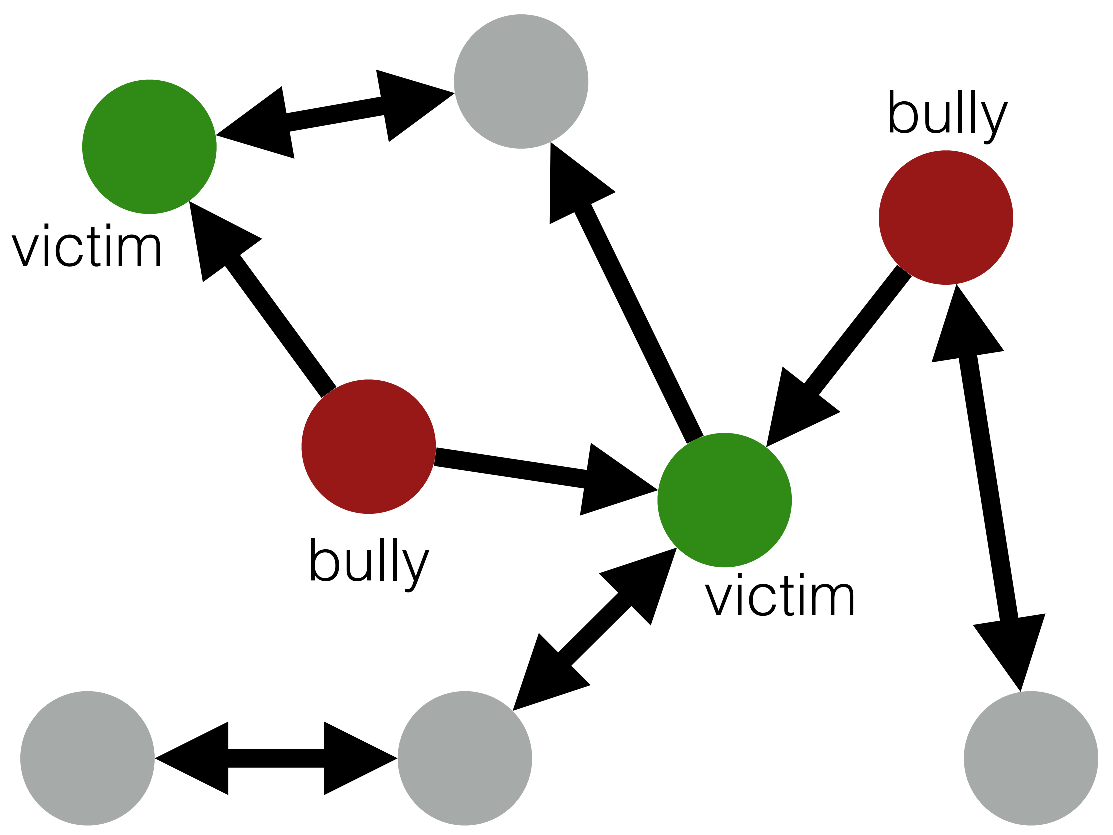
Challenges for Detecting Cyberbullying with Machine Learning

- Social structure is important
- Need scalable algorithms for massive data
- Language is changing:
  - New slang is frequently introduced or old slang becomes outdated
- Annotation:
  - Needs significant consideration of social context



Challenges for Detecting Cyberbullying with Machine Learning

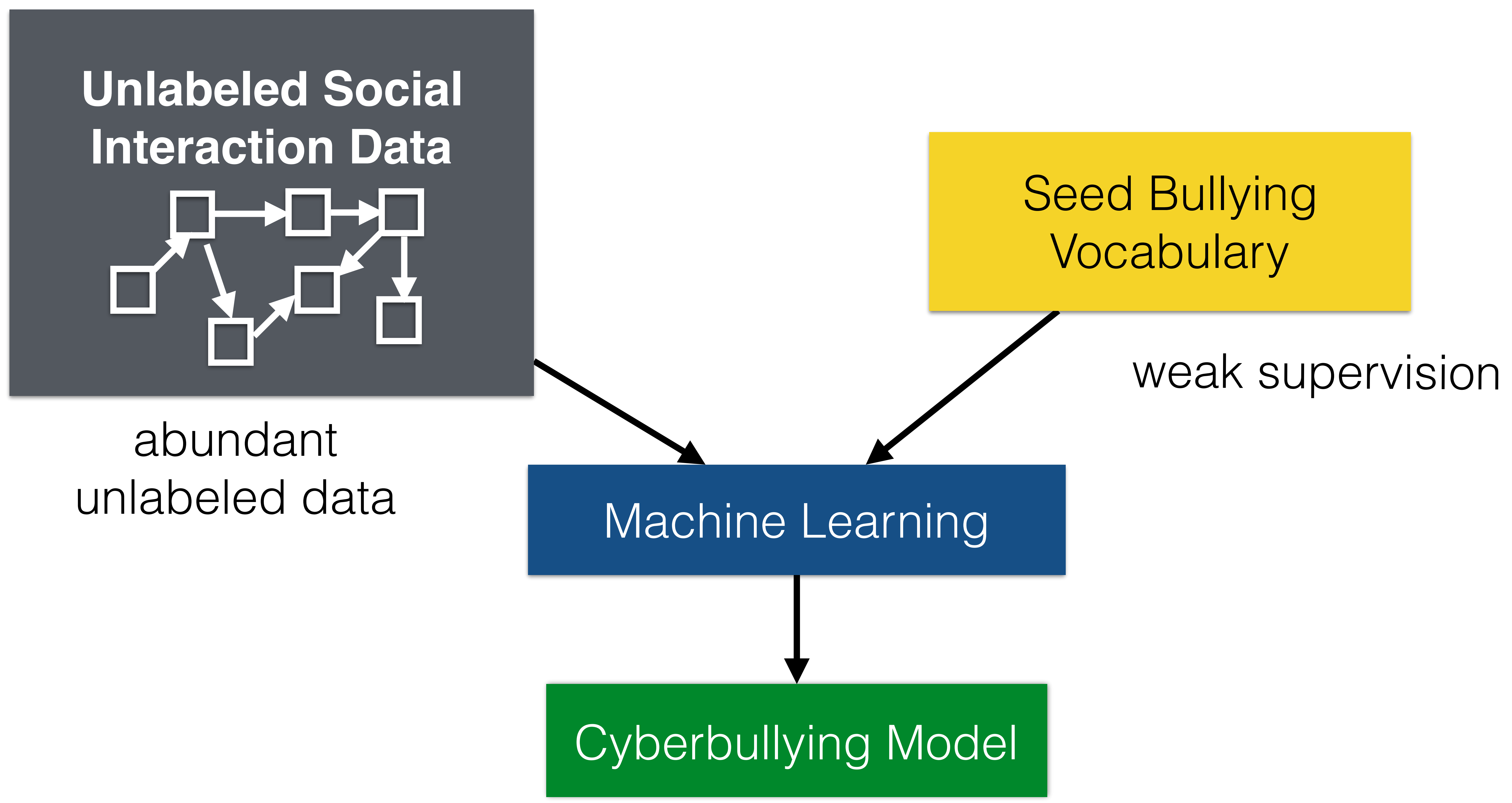
- Social structure is important
- Need scalable algorithms for massive data
- Language is changing:
  - New slang is frequently introduced or old slang becomes outdated
- Annotation:
  - Needs significant consideration of social context
  - Costs add up for a large-scale data



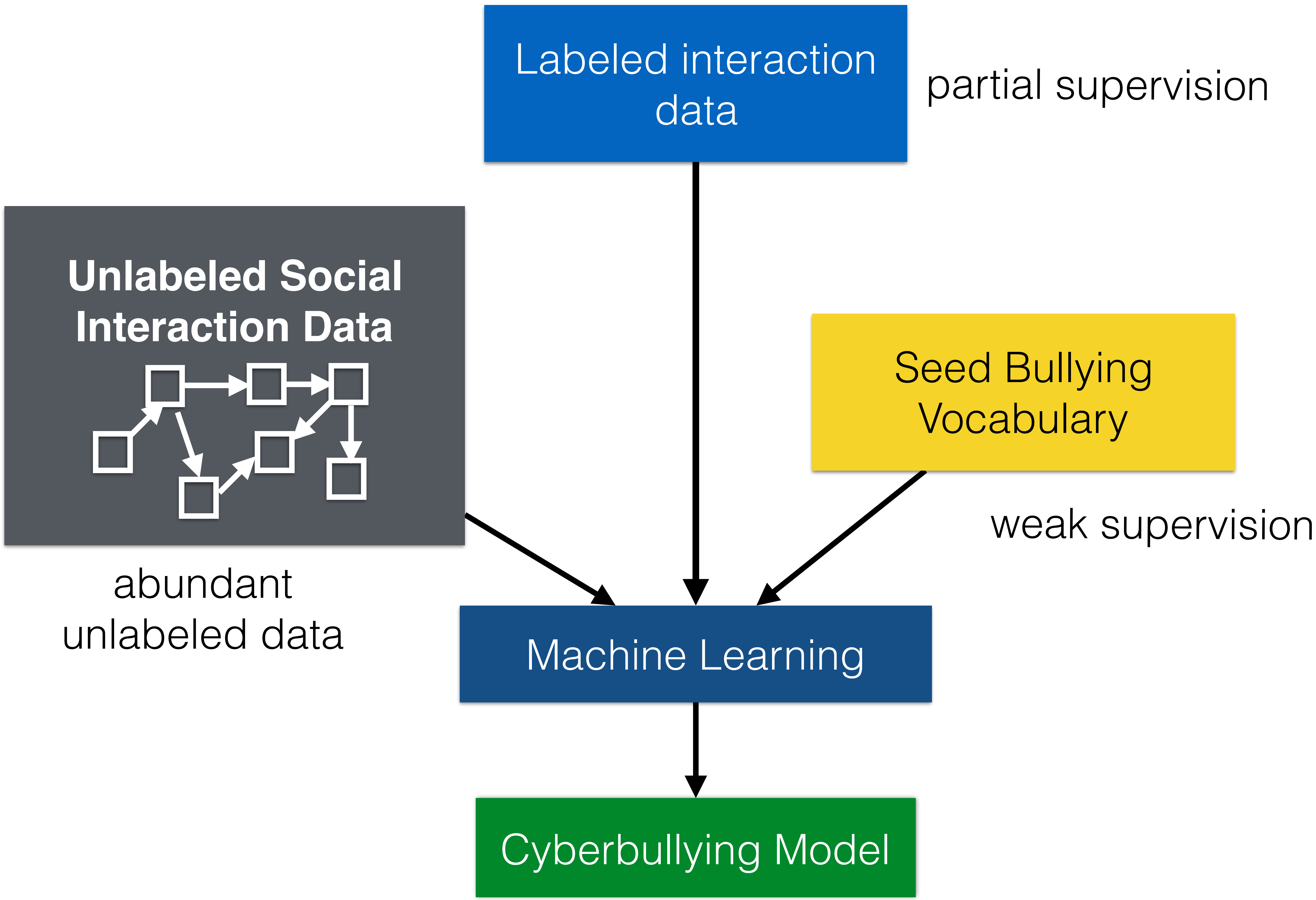
-2-

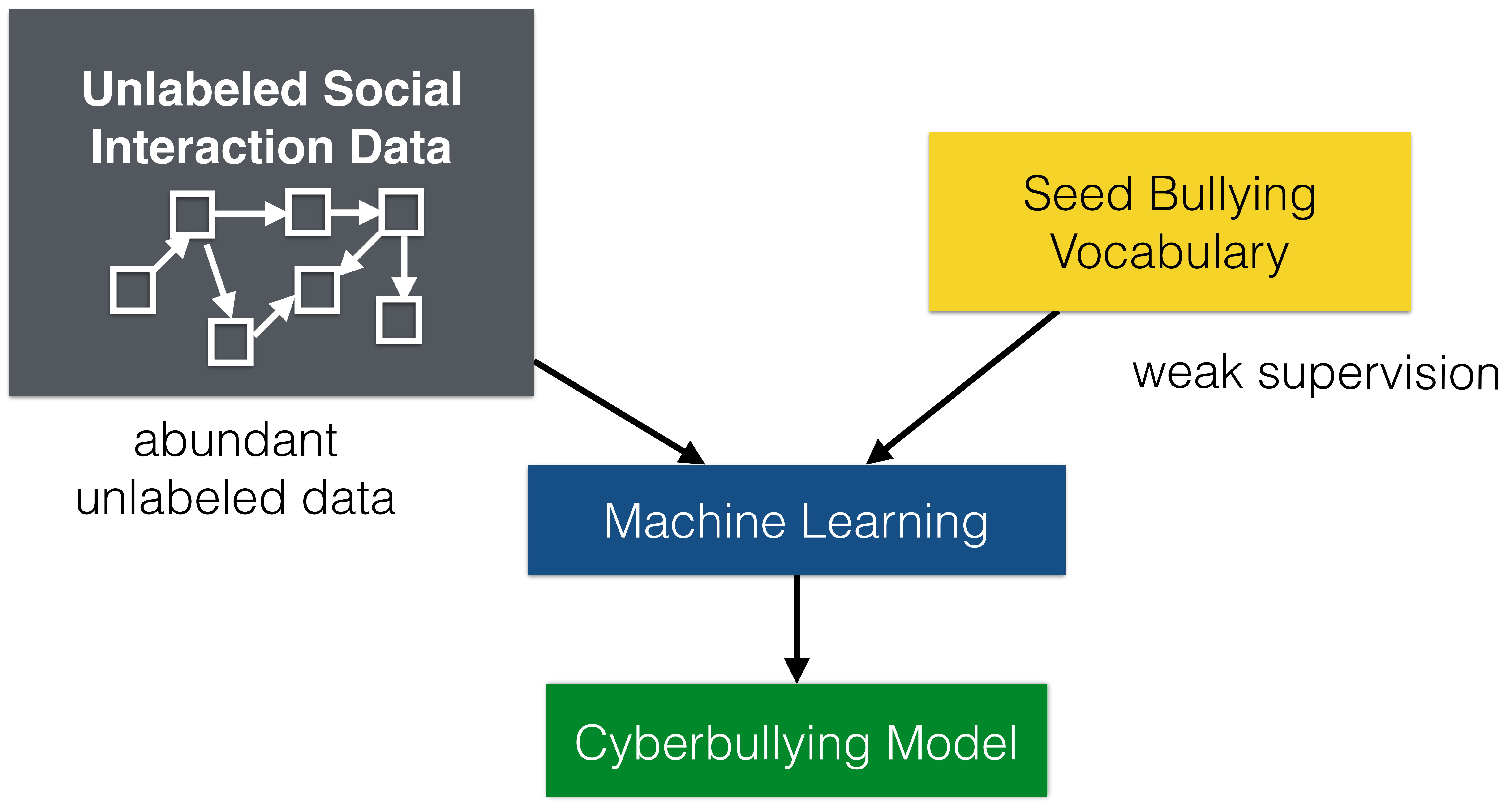
Participant-Vocabulary Consistency

Weakly supervised learning for Cyberbullying Detection



Unlabeled Social Interaction Data
abundant unlabeled data
Seed Bullying Vocabulary
weak supervision
Machine Learning
Cyberbullying Model


Labeled interaction data
partial supervision
Unlabeled Social Interaction Data
Seed Bullying Vocabulary
weak supervision
abundant unlabeled data
Machine Learning
Cyberbullying Model


Unlabeled Social Interaction Data
abundant unlabeled data
Seed Bullying Vocabulary
weak supervision
Machine Learning
Cyberbullying Model

Participant-Vocabulary Consistency Model

Participant-Vocabulary Consistency Model

- Each user has a **bully score** and a **victim score**

Participant-Vocabulary Consistency Model

- Each user has a **bully score** and a **victim score**
- Each n-gram has a **vocabulary score**

Participant-Vocabulary Consistency Model

- Each user has a **bully score** and a **victim score**
- Each n-gram has a **vocabulary score**
- Expert provides seed set of n-grams that we fix to have harassment score 1.0

Participant-Vocabulary Consistency Model

- Each user has a **bully score** and a **victim score**
- Each n-gram has a **vocabulary score**
- Expert provides seed set of n-grams that we fix to have harassment score 1.0

regularizer for all messages vocabulary score of word

$$\min_{\mathbf{b},\mathbf{v},\mathbf{w}} \; \frac{\lambda}{2}\left(||\mathbf{b}||^2 + ||\mathbf{v}||^2 + ||\mathbf{w}||^2\right) + \frac{1}{2}\sum_{m\in M}\left(\sum_{k:w_k\in f(m)}\left(b_{s(m)} + v_{r(m)} - w_k\right)^2\right)$$

$$\text{s.t.} \quad \mathbf{w}_k = 1.0 \text{ for } k \in S$$

bully score of sender victim score of receiver

expert-provided seed set for words in message regularizer

for all messages

vocabulary score of word

$$\min_{\mathbf{b},\mathbf{v},\mathbf{w}} \quad \frac{\lambda}{2}\left(||\mathbf{b}||^2 + ||\mathbf{v}||^2 + ||\mathbf{w}||^2\right) + \frac{1}{2}\sum_{m \in M}\left(\sum_{k:w_k \in f(m)} \left(b_{s(m)} + v_{r(m)} - w_k\right)^2\right)$$

$$\text{s.t.} \quad \mathbf{w}_k = 1.0 \text{ for } k \in S$$

bully score of sender

victim score of receiver

expert-provided seed set

for words in message

$$\min_{\mathbf{b},\mathbf{v},\mathbf{w}} \quad \frac{\lambda}{2}\overbrace{\left(||\mathbf{b}||^2 + ||\mathbf{v}||^2 + ||\mathbf{w}||^2\right)}^{\text{regularizer}} + \frac{1}{2}\overbrace{\sum_{m \in M}}^{\text{for all messages}} \left(\underbrace{\sum_{k:w_k \in f(m)}}_{\text{for words in message}} \left(\underbrace{b_{s(m)}}_{\text{bully score of sender}} + \underbrace{v_{r(m)}}_{\text{victim score of receiver}} - \overbrace{w_k}^{\text{vocabulary score of word}}\right)^2 \right)$$

$$\text{s.t.} \quad \mathbf{w}_k = 1.0 \text{ for } k \in \underbrace{S}_{\text{expert-provided seed set}}$$

Alternating Least Squares

- Objective J(**b**,**v**,**w**,λ) isn't jointly convex
- Alternating least squares:
  - Fix all but one parameter vector at a time
  - Optimize each parameter vector in isolation (closed form)
  - Run until convergence

Participant-Vocabulary Consistency Algorithm

Algorithm 1 Participant-Vocabulary Consistency using Alternating Least-Squares

procedure ParticipantVocabConsistency(b, v, w, λ)
  $\mathbf{b} \leftarrow (0.1, 0.1, 0.1, ..., 0.1)$ ▷ initialize bully scores
  $\mathbf{v} \leftarrow (0.1, 0.1, 0.1, ..., 0.1)$ ▷ initialize victim scores
  $\mathbf{w} \leftarrow (0.1, 0.1, 0.1, ..., 0.1)$ ▷ initialize n-gram scores
  score $\leftarrow J(\mathbf{b}, \mathbf{v}, \mathbf{w}, \lambda)$ ▷ compute objective
  while True **do**
    $\mathbf{b} \leftarrow \left[\arg\min_{b_i} J\right]_{i=1}^{n}$ ▷ update **b**
    $\mathbf{v} = \left[\arg\min_{v_i} J\right]_{i=1}^{n}$ ▷ update **v**
    $\mathbf{w} = \left[\arg\min_{w_k} J\right]_{k=1}^{|\mathbf{w}|}$ ▷ update **w**
    newScore $\leftarrow J(\mathbf{b}, \mathbf{v}, \mathbf{w}, \lambda)$ ▷ new objective
    diff $\leftarrow$ score - newScore
    if diff $< \epsilon$ **then** ▷ convergence tolerance
      break
    score $\leftarrow$ newScore
  return (b, v, w) ▷ returns the final bully, victim score of users and the score of words

Experiments

	# Users after preprocessing	# Messages after preprocessing
Ask.fm	260,800	2,863,801
Instagram	3,829,756	9,828,760
Twitter	180,355	296,308

Instagram and ask.fm data from [Hosseinmardi et al., CoRR '14]

noswearing.com	3,461 offensive unigrams and bigrams

Baseline Algorithms

Baseline Algorithms

- **Seed words**: use only seed words as bullying vocabulary

Baseline Algorithms

- **Seed words**: use only seed words as bullying vocabulary
- **Co-occurrence**: add words to bullying vocab. if they appear in messages with seed words

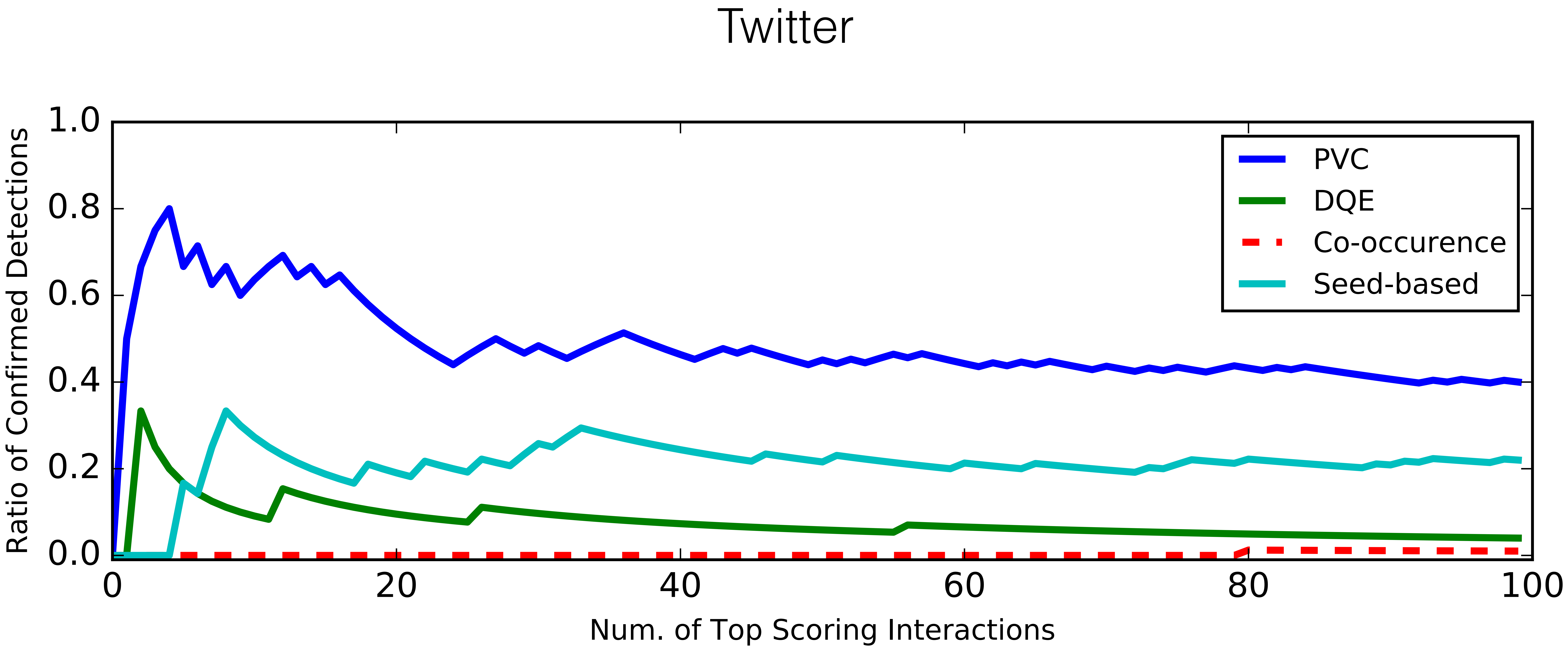
Baseline Algorithms

- **Seed words**: use only seed words as bullying vocabulary
- **Co-occurrence**: add words to bullying vocab. if they appear in messages with seed words
- **Dynamic query expansion** (DQE) [Ramakrishnan, KDD '14]
    1. For every word that co-occurs with current bullying vocabulary, compute its *document frequency*
    2. Add the N highest-scoring keywords to vocabulary
    3. Repeat until convergence

Post-Hoc Analysis: Conversations

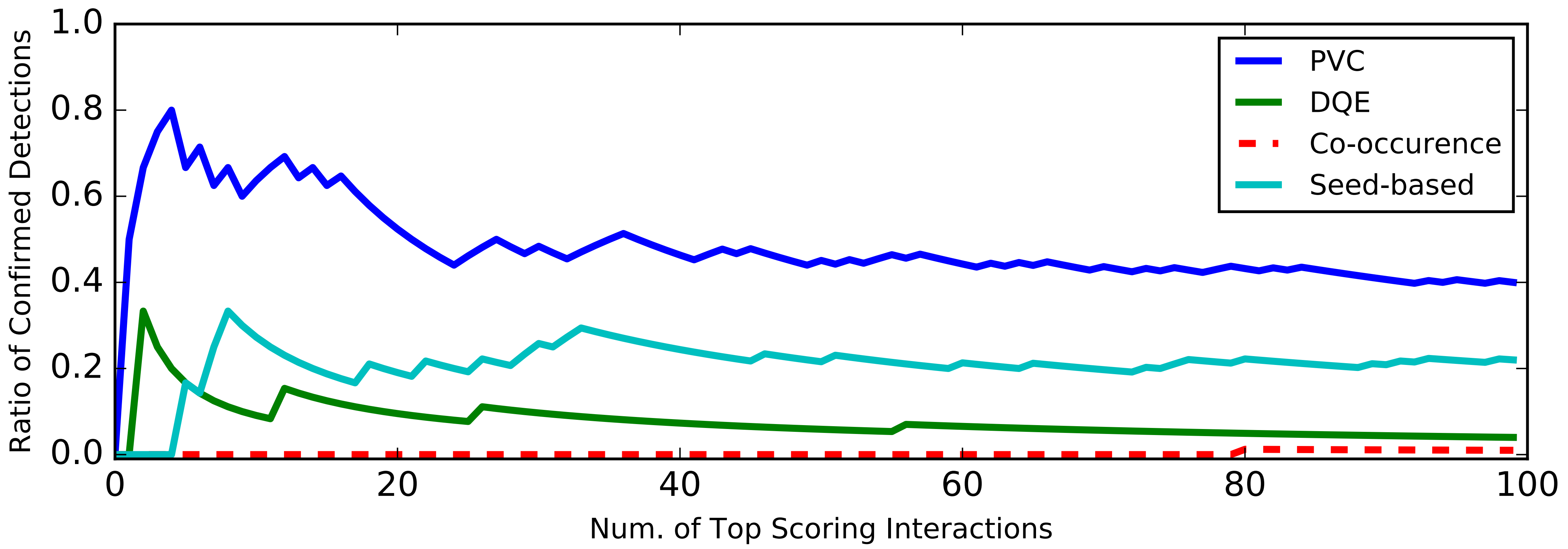
- Each method: extract 100 conversations most likely to be bullying
- Three annotators rate as “yes”, “no”, or “uncertain”
- Consider each conversation with majority yes votes relevant; compute precision@k

Post-Hoc Analysis: Conversations

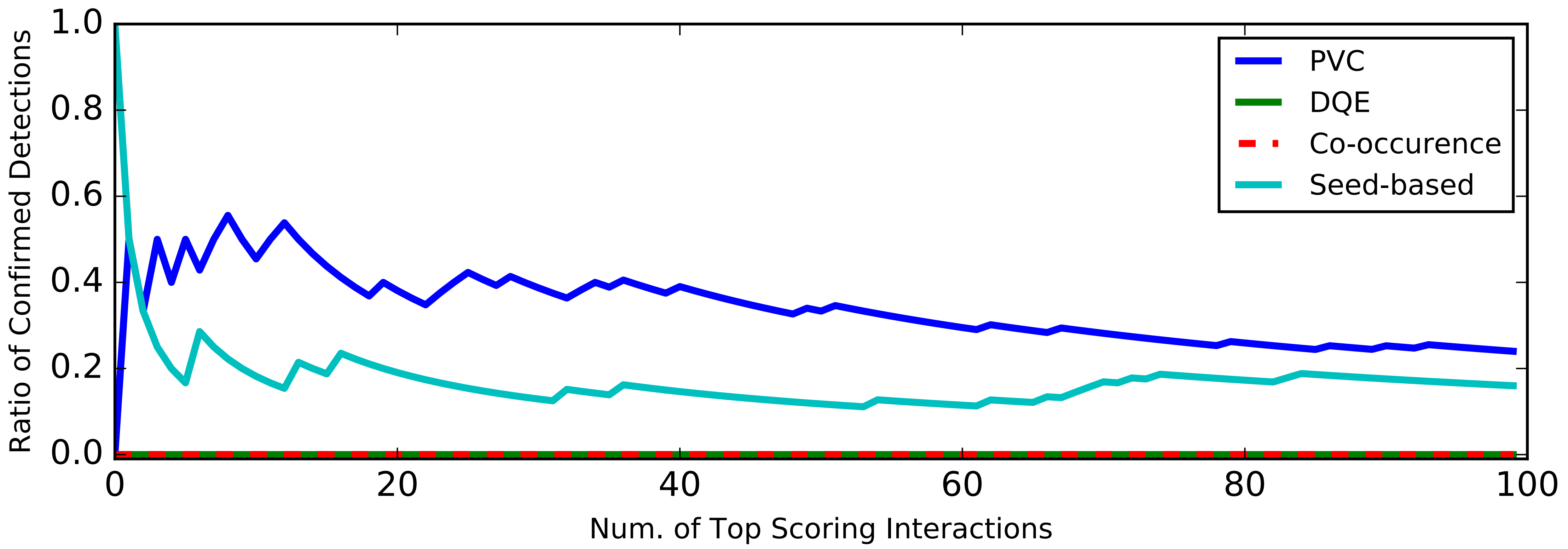


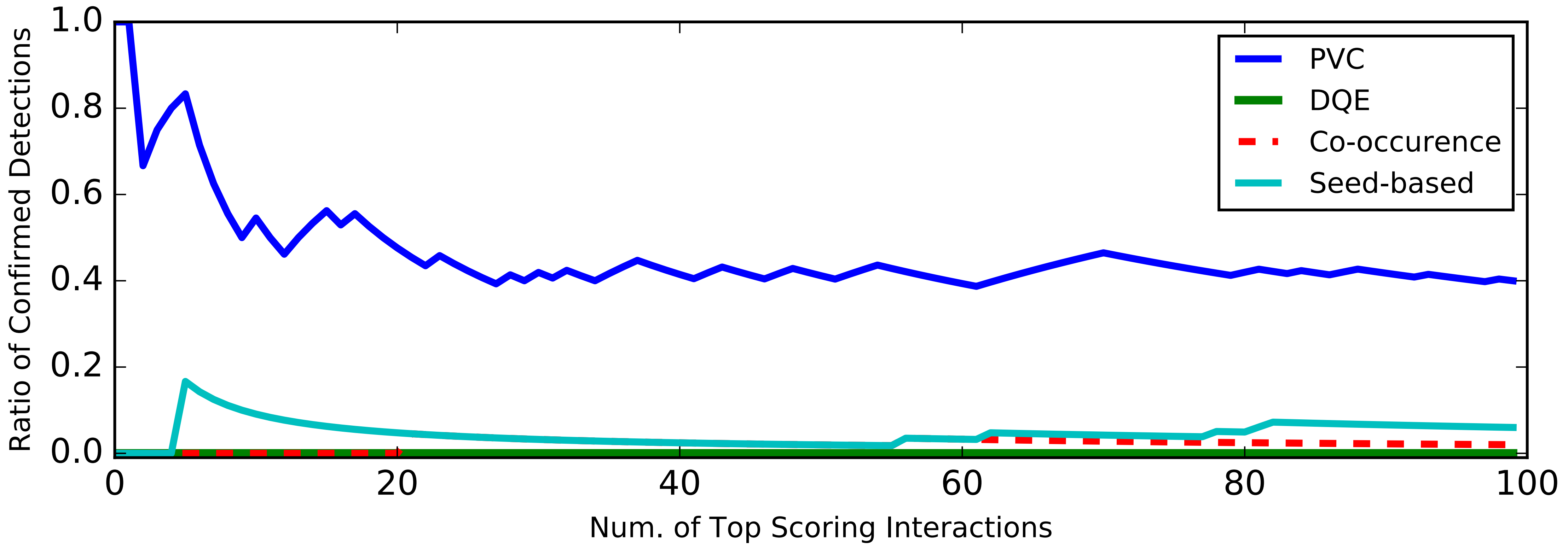
Post-Hoc Analysis: Conversations

Twitter



Instagram



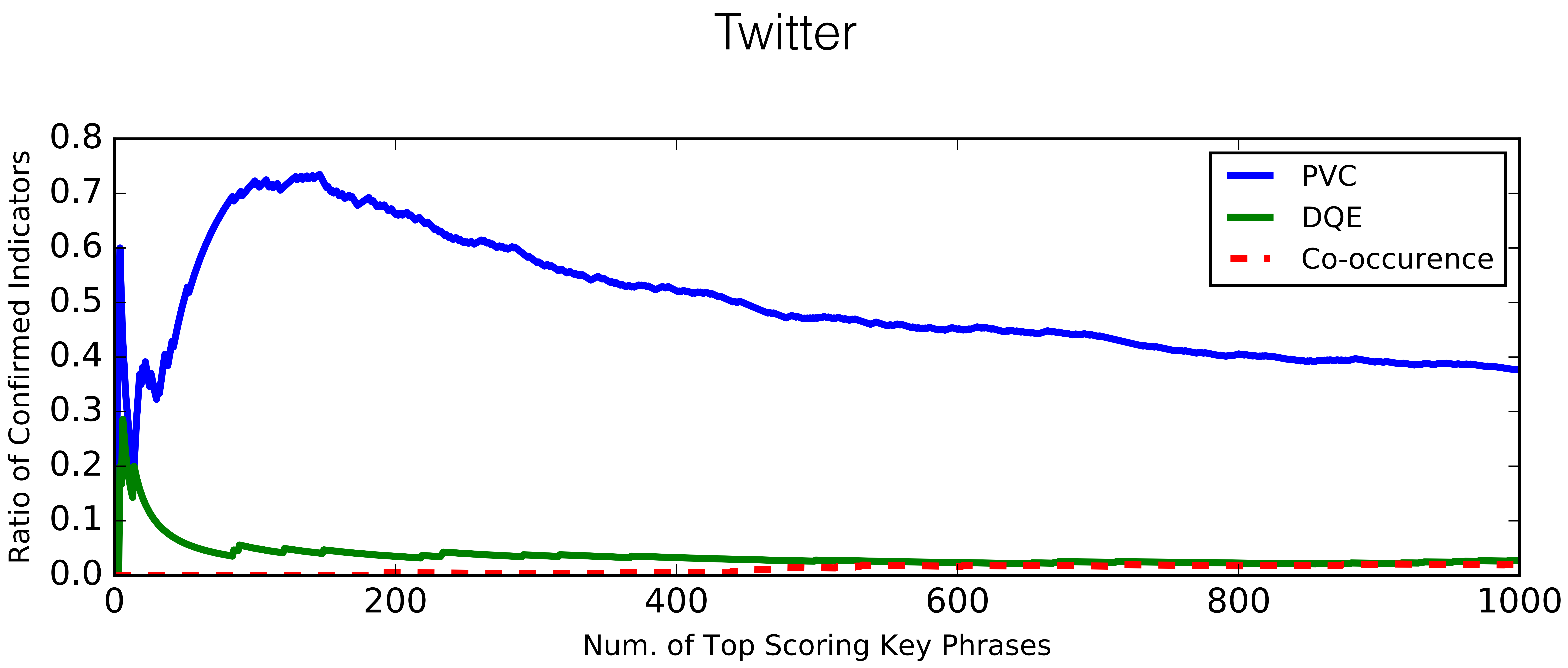


Ask.fm

Post-Hoc Analysis: Key Phrases

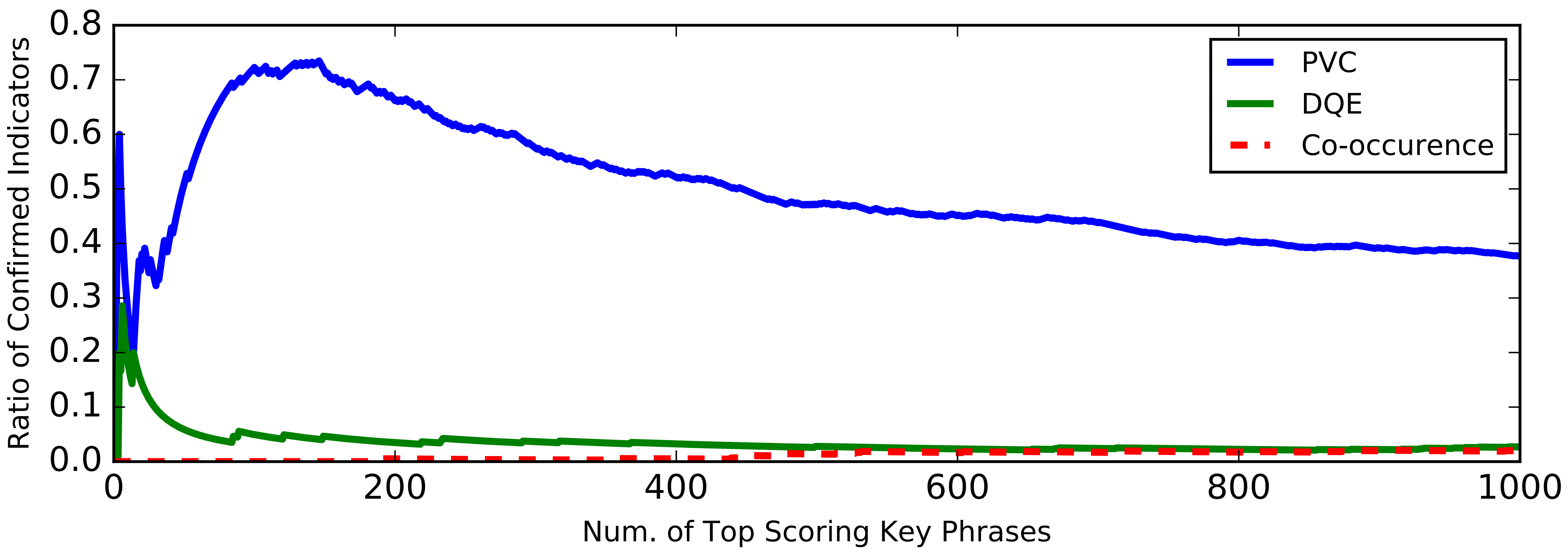
- Each method: 1000 strongest key phrase indicators
- Three annotators rate as “yes”, “no”, or “uncertain”
- Consider each key phrase with majority yes votes relevant; compute precision@k

Post-Hoc Analysis: Key Phrases

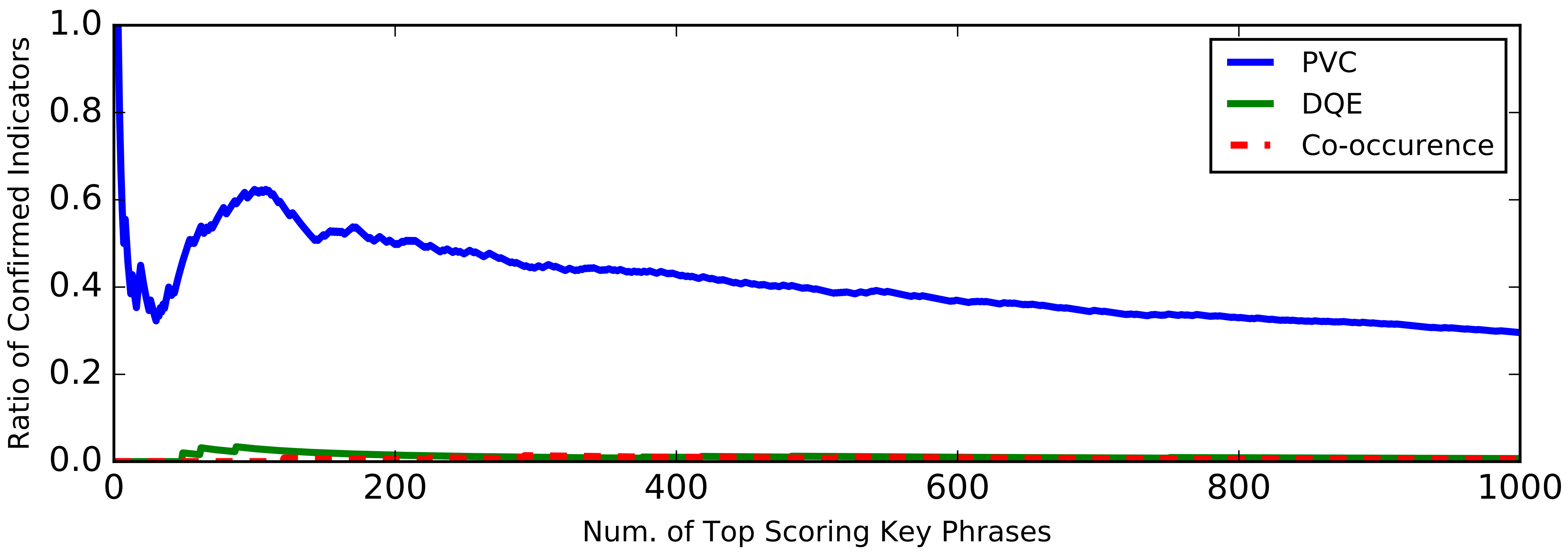


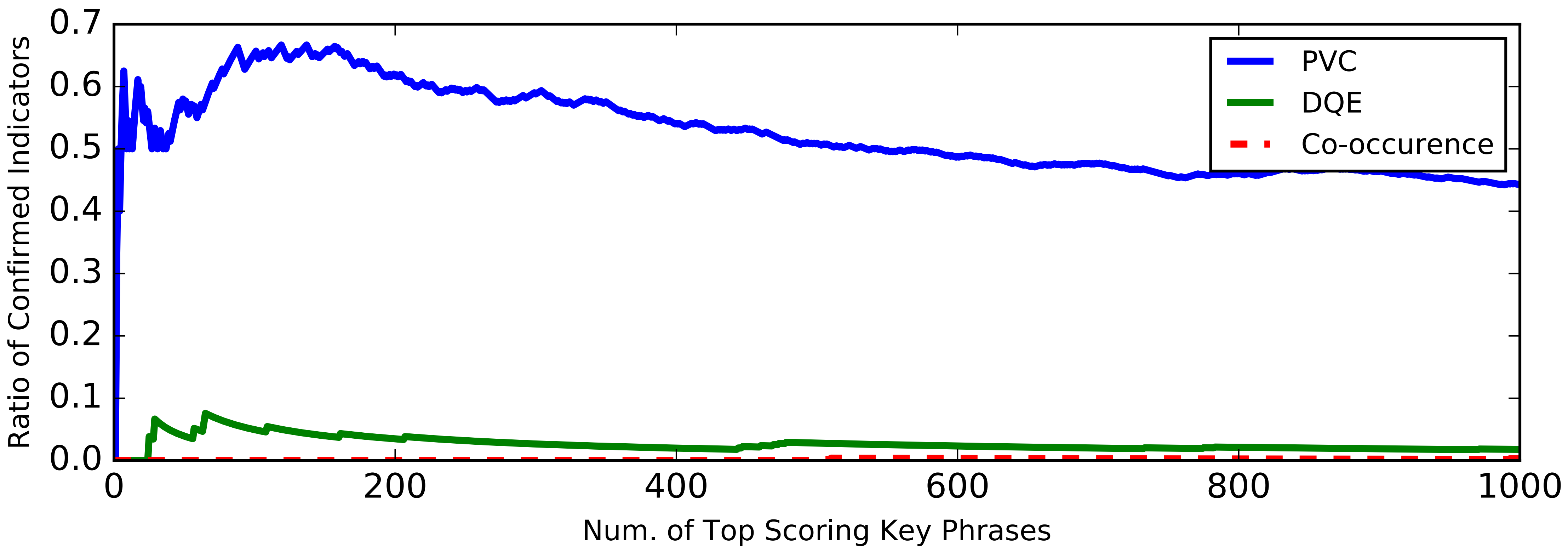
Post-Hoc Analysis: Key Phrases

Twitter



Instagram





Ask.fm

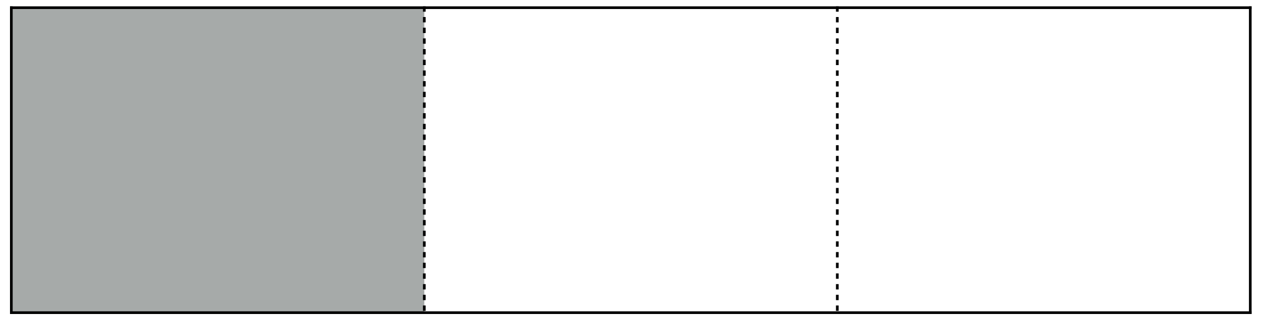
Post-Hoc Analysis

Twitter

Method	Detected Bullying Words Color-Coded by Annotation: Bullying, Likely Bullying, Uncertain, Not Bullying.
PVC	singlemost biggest, singlemost, delusional prick, existent *ss, biggest jerk, karma bites, hope karma, jerk milly, rock freestyle, jay jerk, worldpremiere, existent, milly rock, milly, freestyle, *ss b*tch, d*ck *ss, *ss hoe, b*tch *ss, adore black, c*mming f*ck, tgurl, tgurl sl*t, black males, rt super, super annoying, sl*t love, bap babyz, love rt, f*ck follow, babyz, jerk *ss, love s*ck, hoe *ss, c*nt *ss, *ss c*nt, stupid *ss, bap, karma, *ss *ss, f*ggot *ss, weak *ss, bad *ss, nasty *ss, lick *ss, d*ck s*cker, wh*re *ss, ugly *ss, s*ck *ss, f*ck *ss,
DQE	don, lol, good, amp, f*ck, love, sh*t, ll, time, people, yeah, ve, man, going, f*cking, head, didn, day, better, free, ya, face, great, hey, best, follow, haha, big, happy, gt, hope, check, gonna, thing, nice, feel, god, work, game, doesn, thought, lmao, life, c*ck, help, lt, play, hate, real, today,
CO	drink sh*tfaced, juuust, sh*tfaced tm4l, tm4l, tm4l br, br directed, subscribe, follow check, music video, check youtube, checkout, generate, comment subscribe, rt checkout, ada, follback, marketing, featured, unlimited, pls favorite, video rob, beats amp, untagged, instrumentals, spying, download free, free beats, absolutely free, amp free, free untagged, submit music, untagged beats, free instrumentals, unlimited cs, creative gt, free exposure, followers likes, music chance, soundcloud followers, spying tool, chakras, whatsapp spying, gaming channel, telepaths, telepaths people, youtube gaming, dir, nightclub, link amp, mana

Experiments: Quantitative Analysis

- Collect offensive words, split into **seed set** and held-out **target words** for evaluation

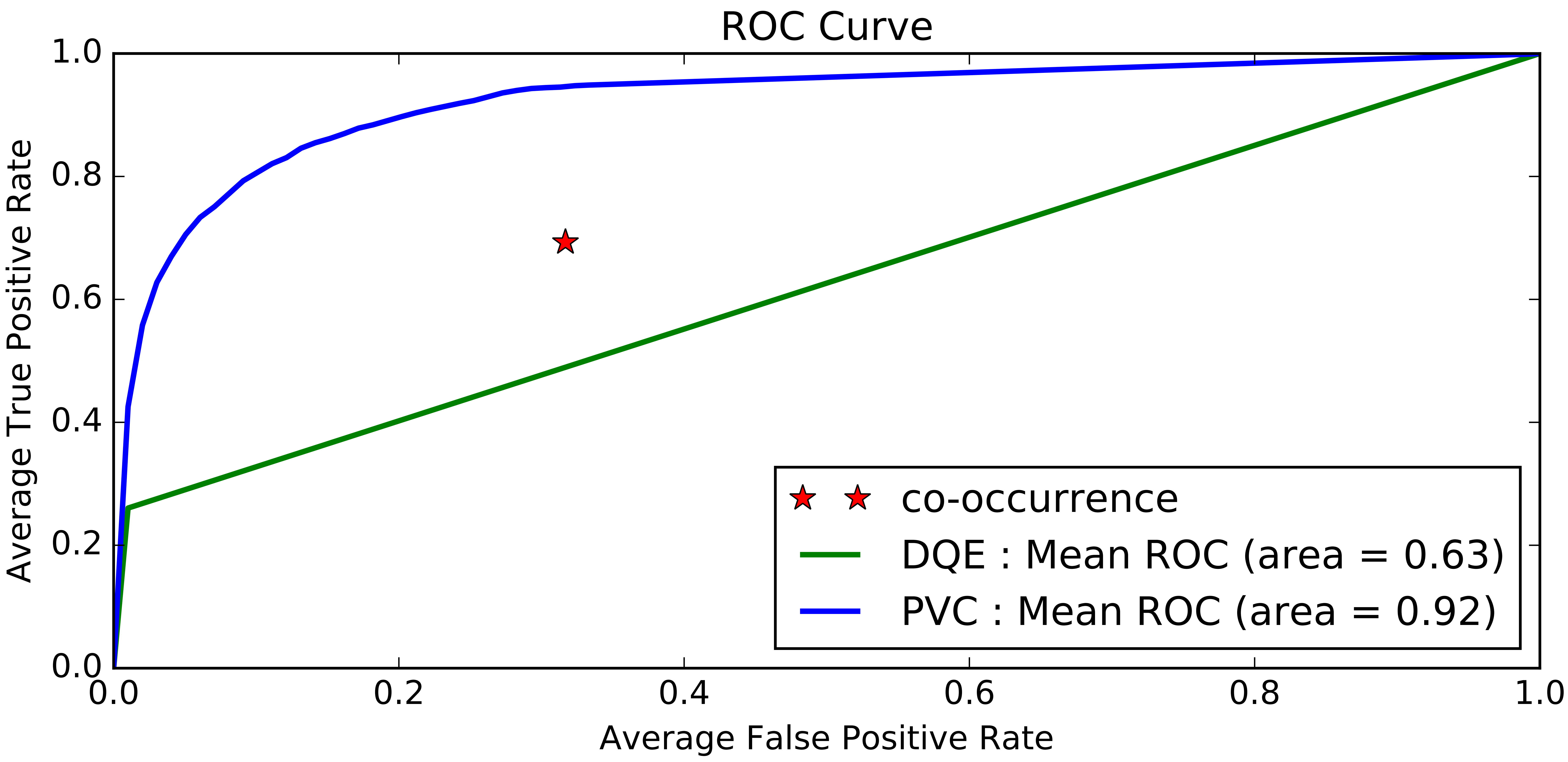


- Evaluation metric: average target-word score. Successful discovery should score target words **higher** than others.

Dataset	Method	Overall Average	Lift (S.D.)
Twitter	PVC	0.001367	**+5.919**
	DQE	1.9663	+0.1276
	CO	0.31698	-0.6811
Ask.fm	PVC	0.0048	**+4.381**
	DQE	1.24e-06	+0.1068
	CO	0.9352	-3.800
Instagram	PVC	0.00706	**+4.1137**
	DQE	5.84e-07	+0.1032
	CO	0.8952	-2.922

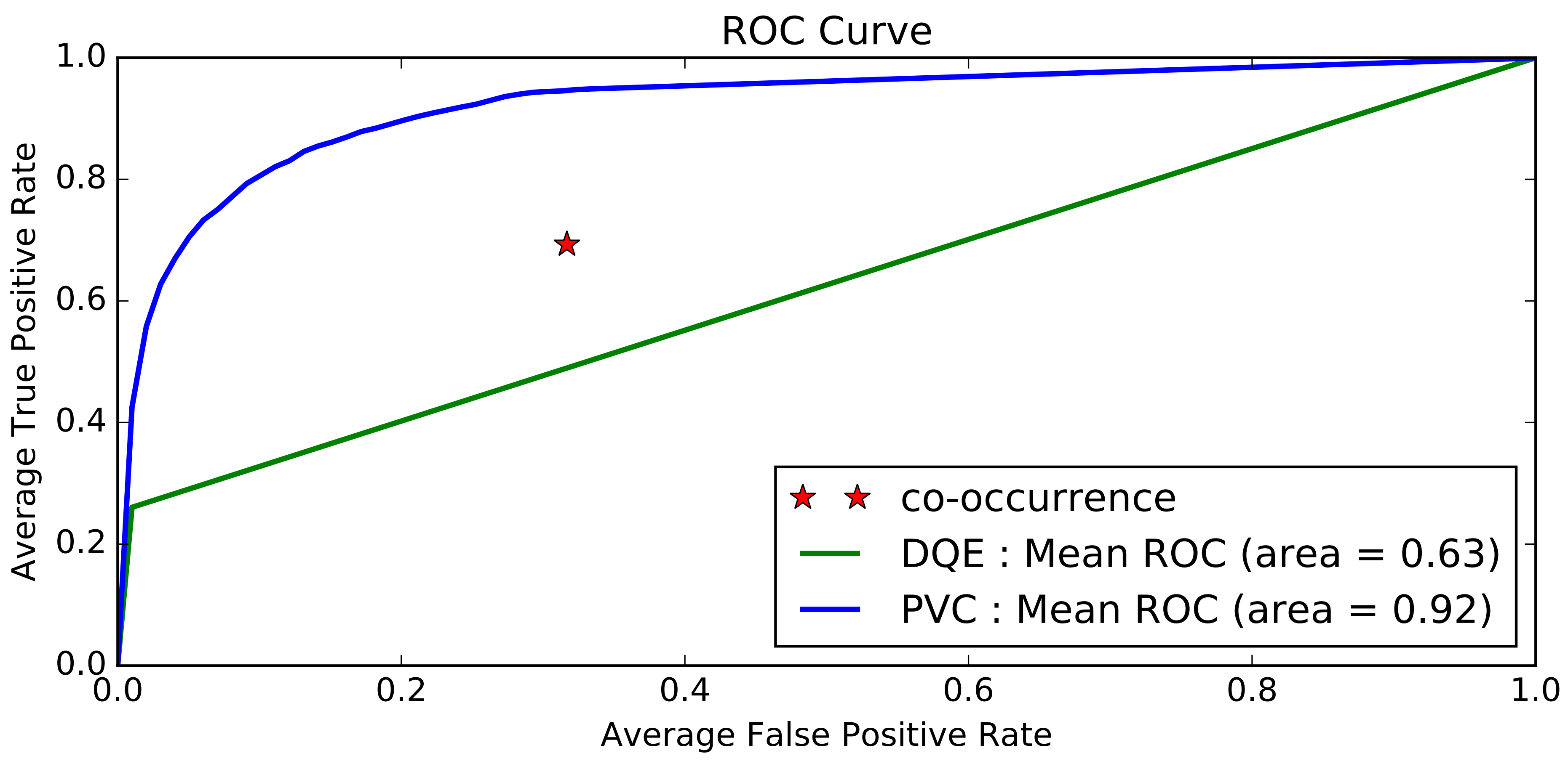
Experiments: Quantitative Analysis

Twitter

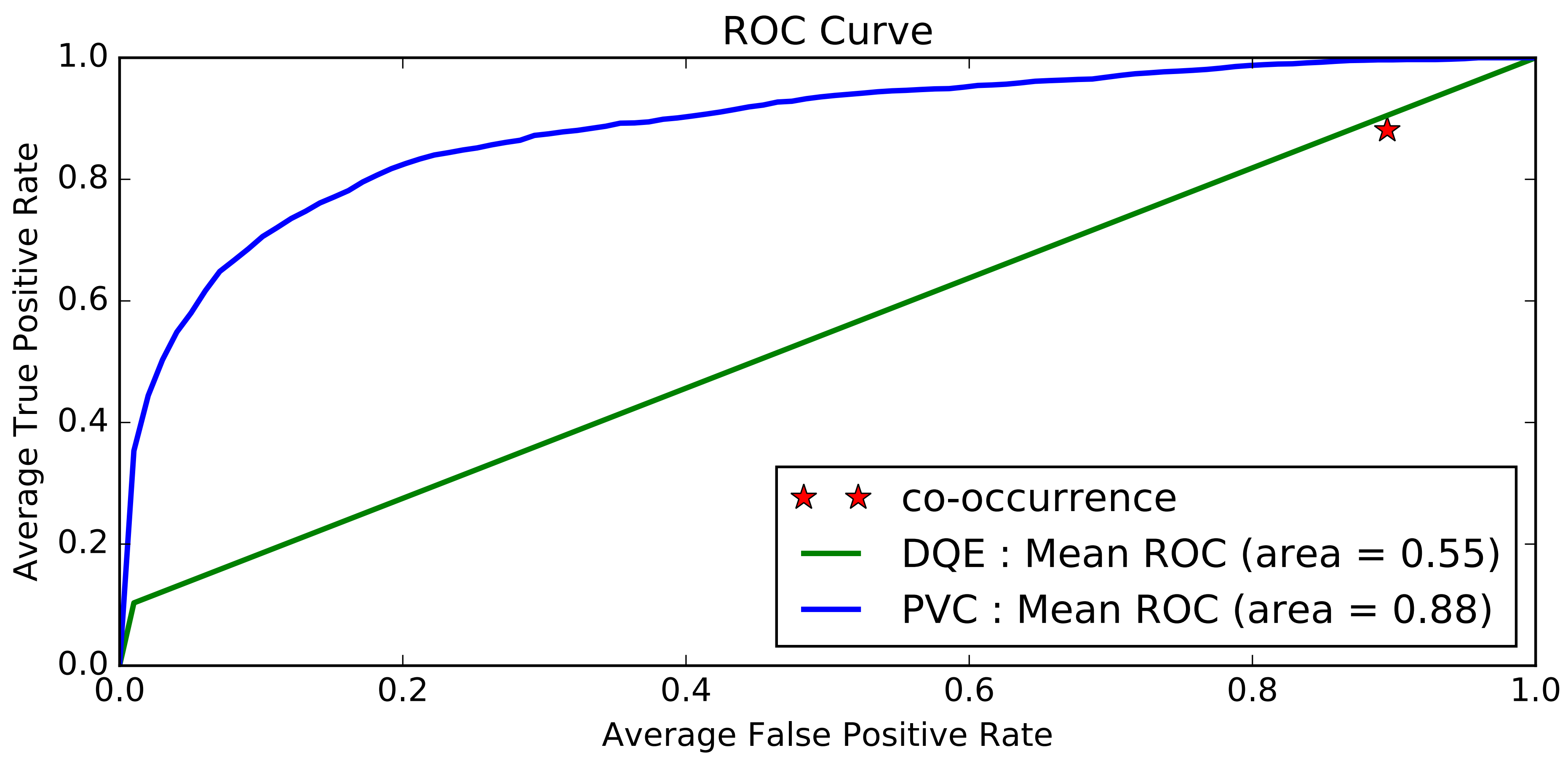


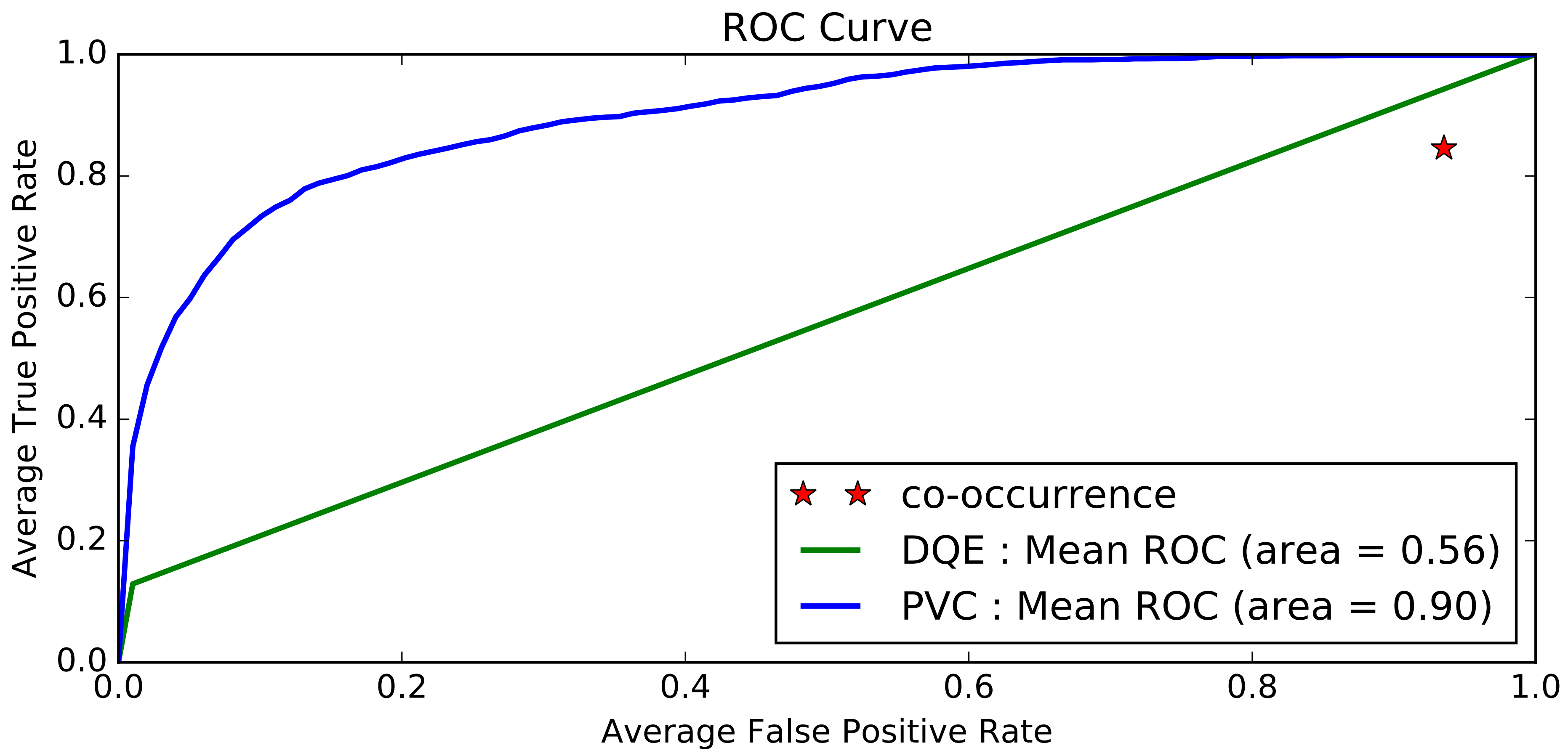
Experiments: Quantitative Analysis

Twitter



Instagram

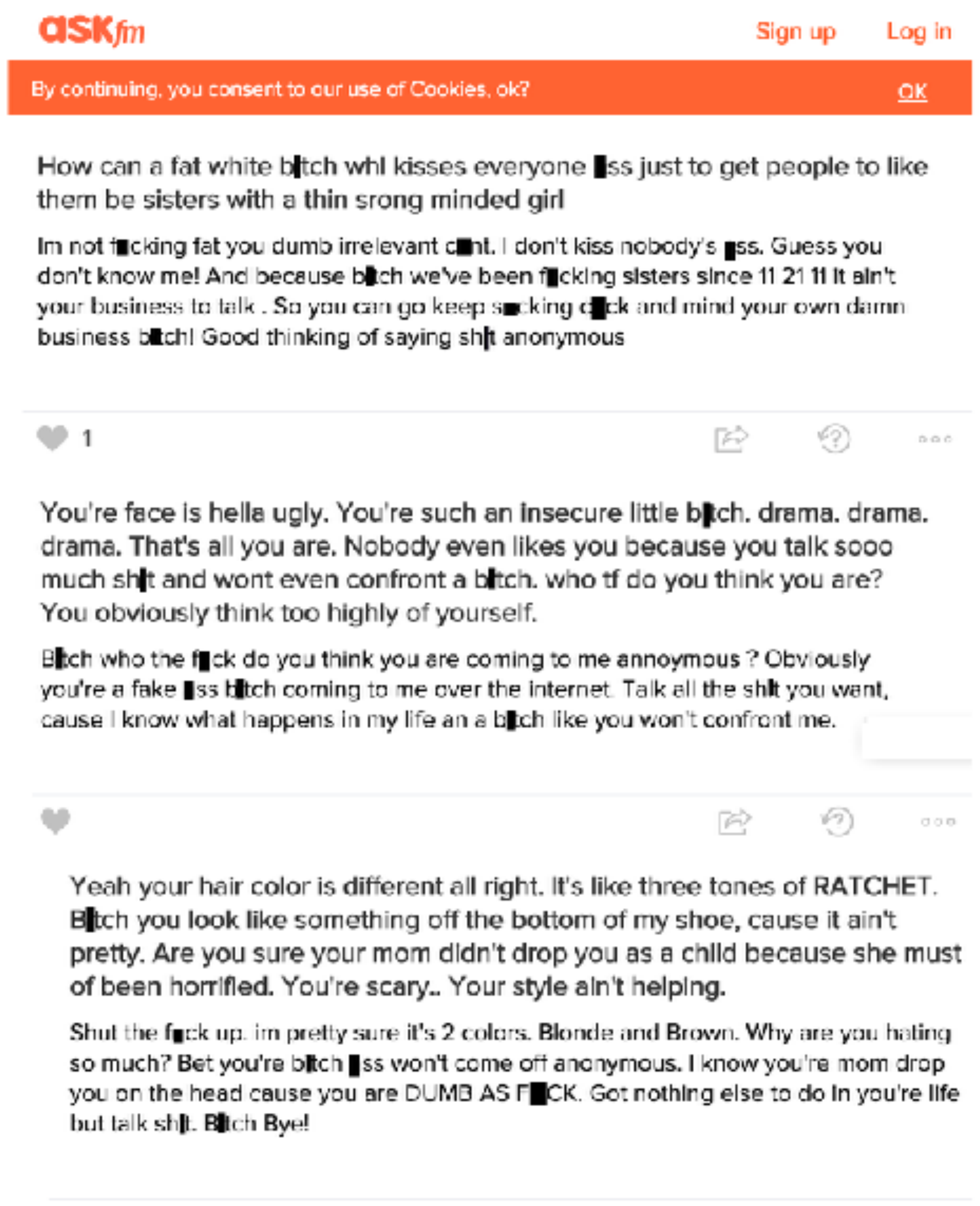




Ask.fm

Experiments: Qualitative Analysis

Example of an Ask.fm conversations of a user PVC gave a high victim score to.



ask fm

Sign up Log in

By continuing, you consent to our use of Cookies, ok? OK

How can a fat white b[redacted]tch whl kisses everyone [redacted]ss just to get people to like them be sisters with a thin srong minded girl

Im not f[redacted]cking fat you dumb irrelevant c[redacted]nt. I don't kiss nobody's [redacted]ss. Guess you don't know me! And because b[redacted]ch we've been f[redacted]cking sisters since 11 21 11 It ain't your business to talk . So you can go keep s[redacted]cking d[redacted]ck and mind your own damn business b[redacted]tch! Good thinking of saying sh[redacted]t anonymous

1

You're face is hella ugly. You're such an insecure little b[redacted]tch. drama. drama. drama. That's all you are. Nobody even likes you because you talk sooo much sh[redacted]t and wont even confront a b[redacted]tch. who tf do you think you are? You obviously think too highly of yourself.

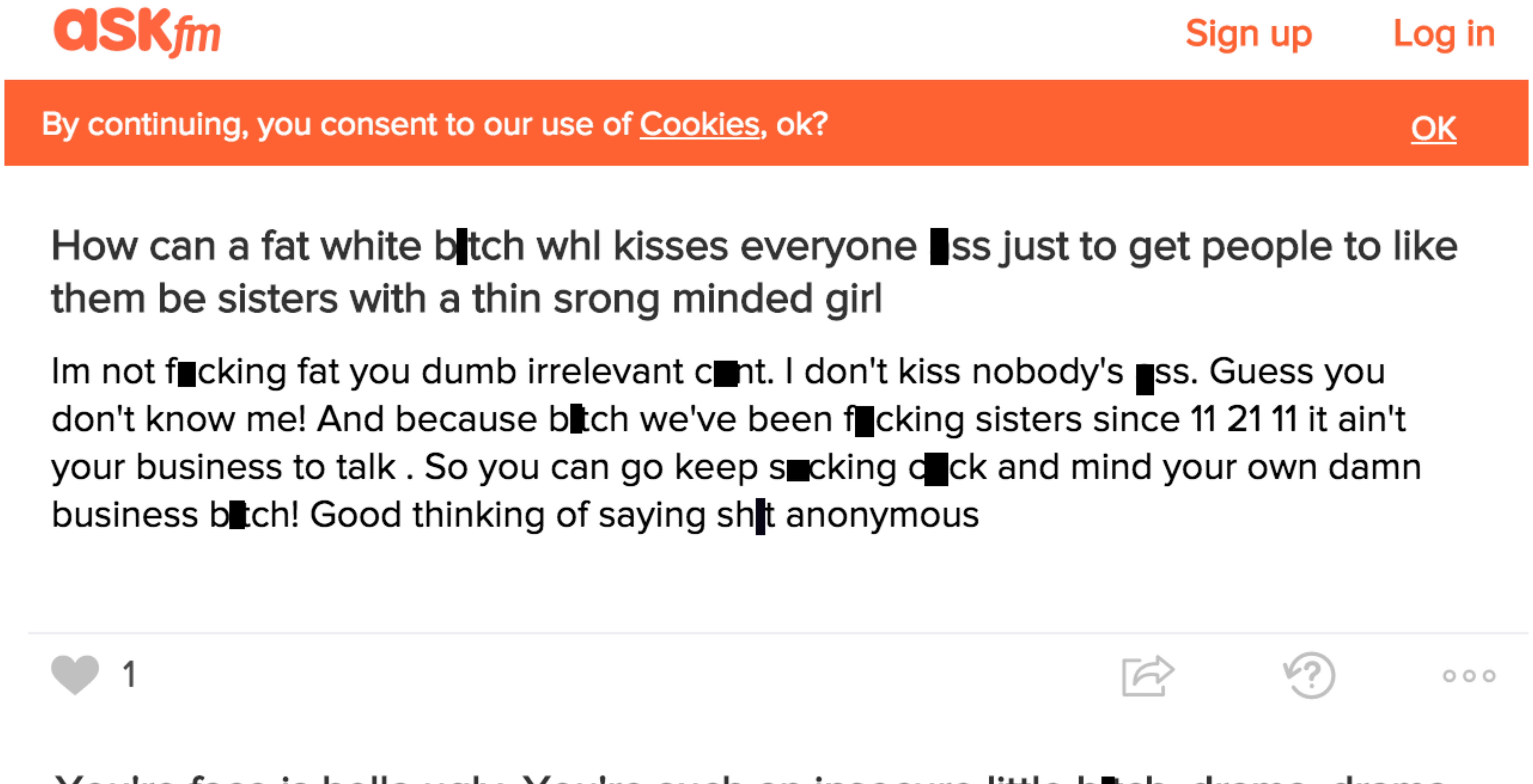
B[redacted]tch who the f[redacted]ck do you think you are coming to me annoymous ? Obviously you're a fake [redacted]ss b[redacted]tch coming to me over the internet. Talk all the sh[redacted]t you want, cause I know what happens in my life an a b[redacted]ch like you won't confront me.

Yeah your hair color is different all right. It's like three tones of RATCHET. B[redacted]tch you look like something off the bottom of my shoe, cause it ain't pretty. Are you sure your mom didn't drop you as a child because she must of been horrified. You're scary.. Your style ain't helping.

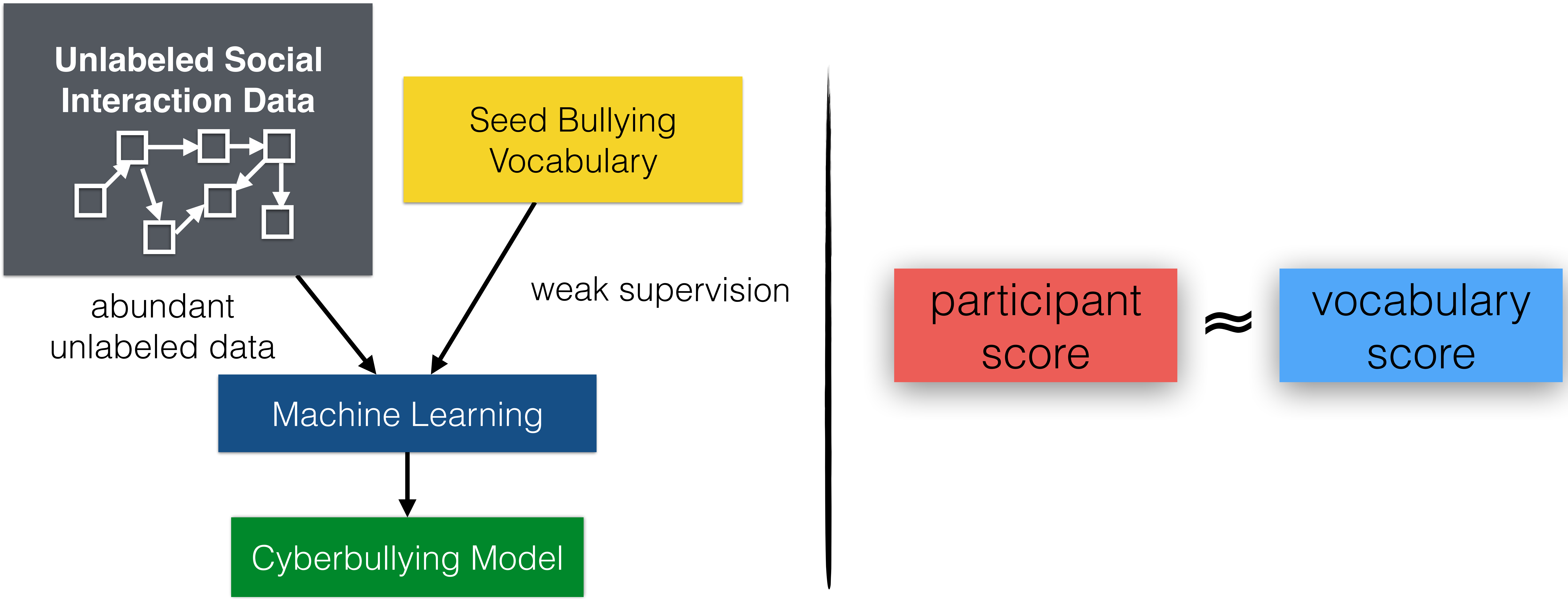
Shut the f[redacted]ck up. im pretty sure it's 2 colors. Blonde and Brown. Why are you hating so much? Bet you're b[redacted]tch [redacted]ss won't come off anonymous. I know you're mom drop you on the head cause you are DUMB AS F[redacted]CK. Got nothing else to do in you're life but talk sh[redacted]t. B[redacted]tch Bye!

Experiments: Qualitative Analysis

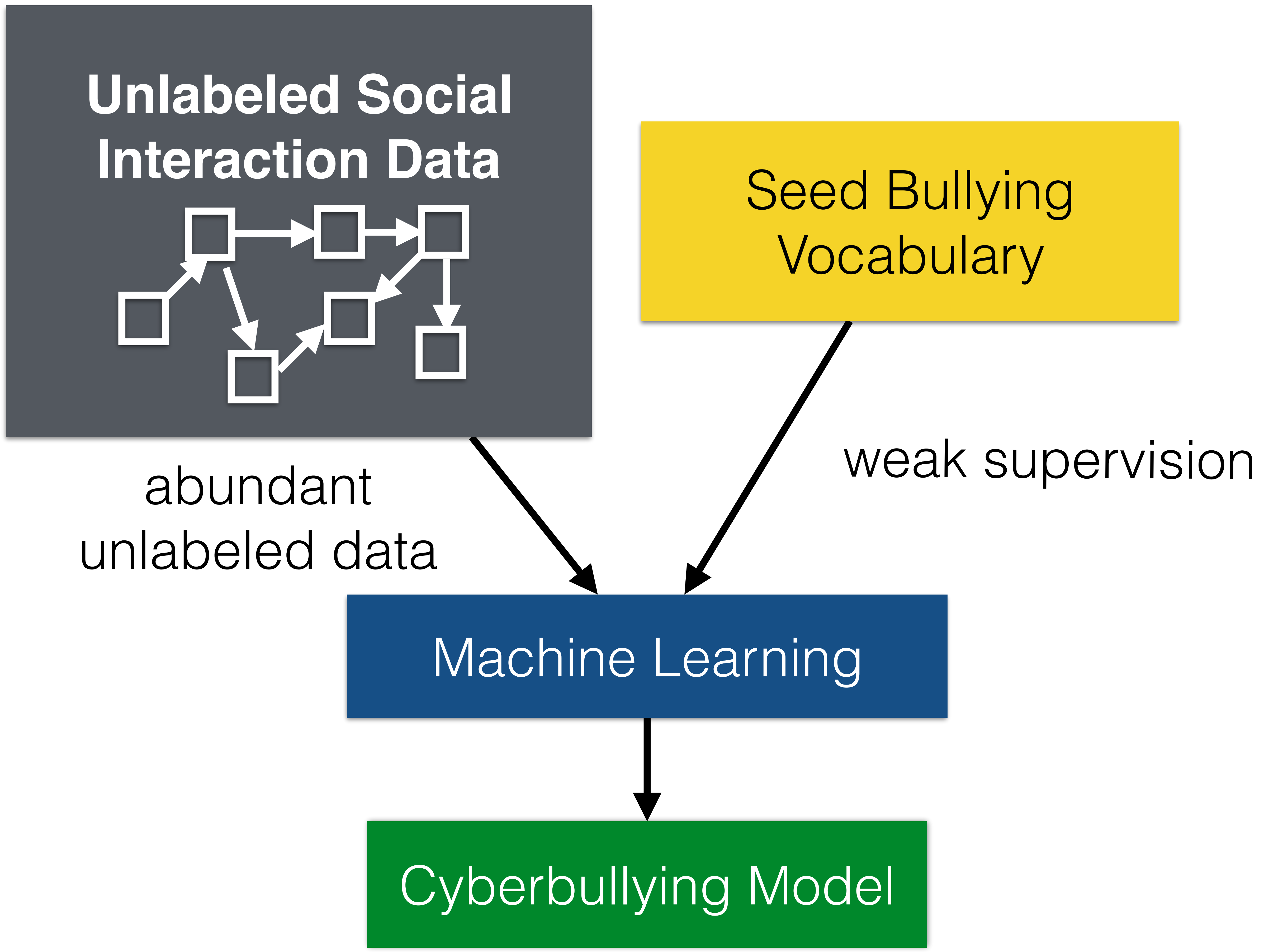
- Example of an Ask.fm conversations of a user PVC gave a high victim score to.

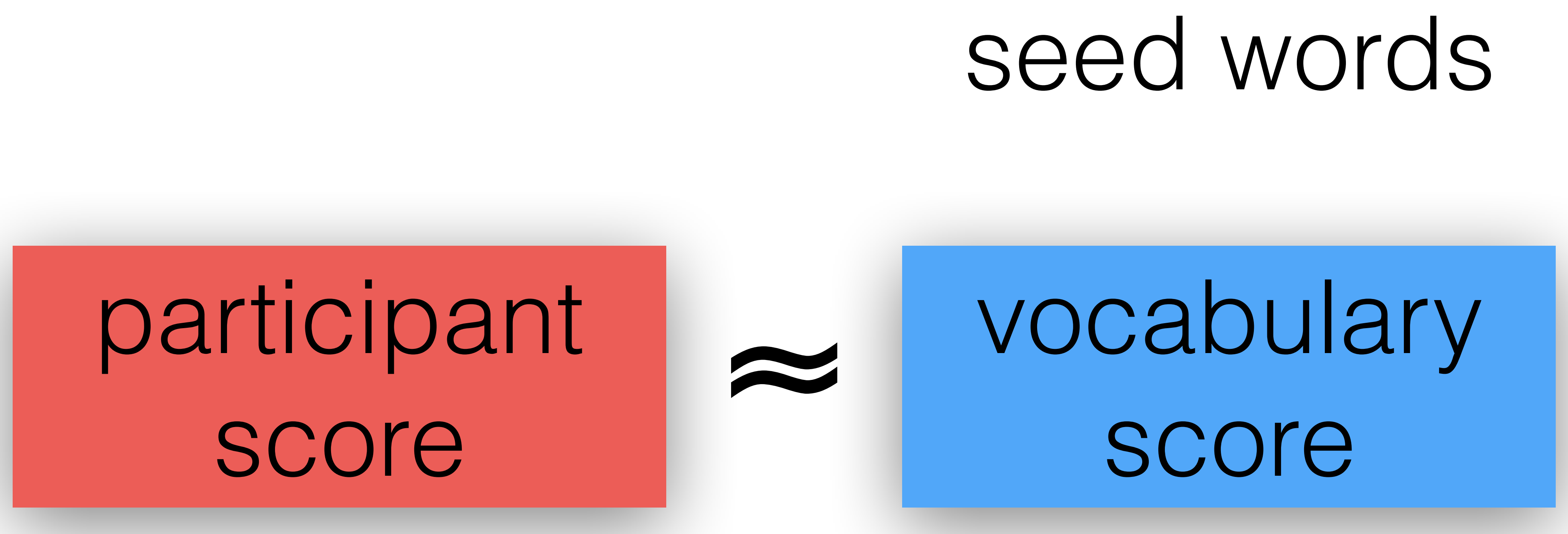


Participant Vocabulary Consistency

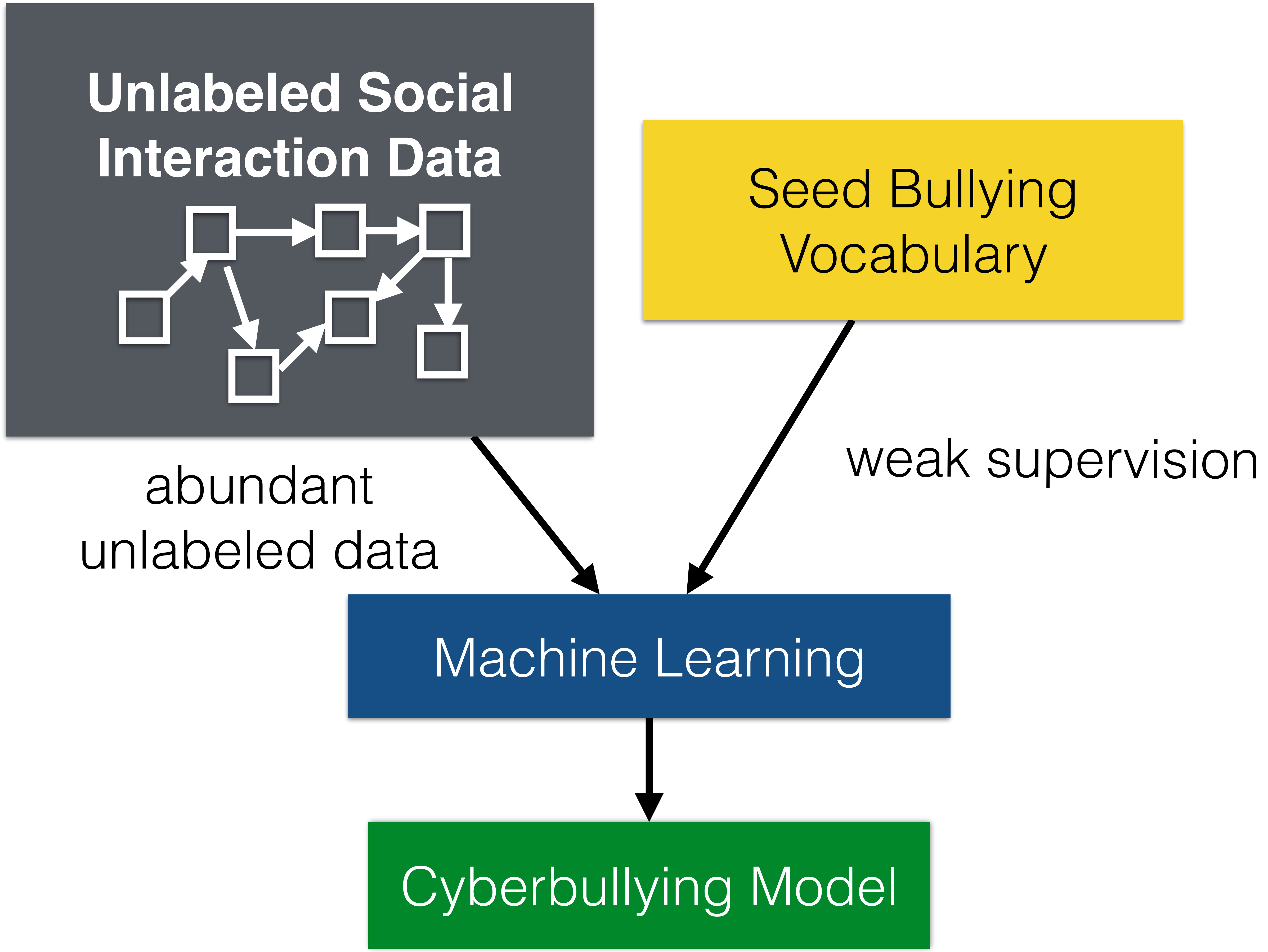


Participant Vocabulary Consistency





Participant Vocabulary Consistency



social structure | seed words

participant score $\approx$ vocabulary score

-3-

Automated Interventions

Key Questions

- Automatic detection will always be noisy. Is it safe to act on uncertainty?
- Even if perfect, what actions prevent or mitigate cyberviolence? What actions actually exacerbate it?
- How will cyberbullies respond to technology meant to thwart their attacks?

Interventions and Consequences

Interventions and Consequences

- Examples: Filtering, advice, human mediation

Interventions and Consequences

- Examples: Filtering, advice, human mediation
- Censorship concerns, false positives, lowered awareness of threats

Interventions and Consequences

- Examples: Filtering, advice, human mediation
- Censorship concerns, false positives, lowered awareness of threats
- Resentment, embarrassment, escalation

Interventions and Consequences

- Examples: Filtering, advice, human mediation
- Censorship concerns, false positives, lowered awareness of threats
- Resentment, embarrassment, escalation
- Trial by fire?

Proposal: A Virtual Social Laboratory

Proposal: A Virtual Social Laboratory

- Online social network with all users role-playing fabricated personas

Proposal: A Virtual Social Laboratory

- Online social network with all users role-playing fabricated personas
- Safe environment to experiment with cybersafety technology

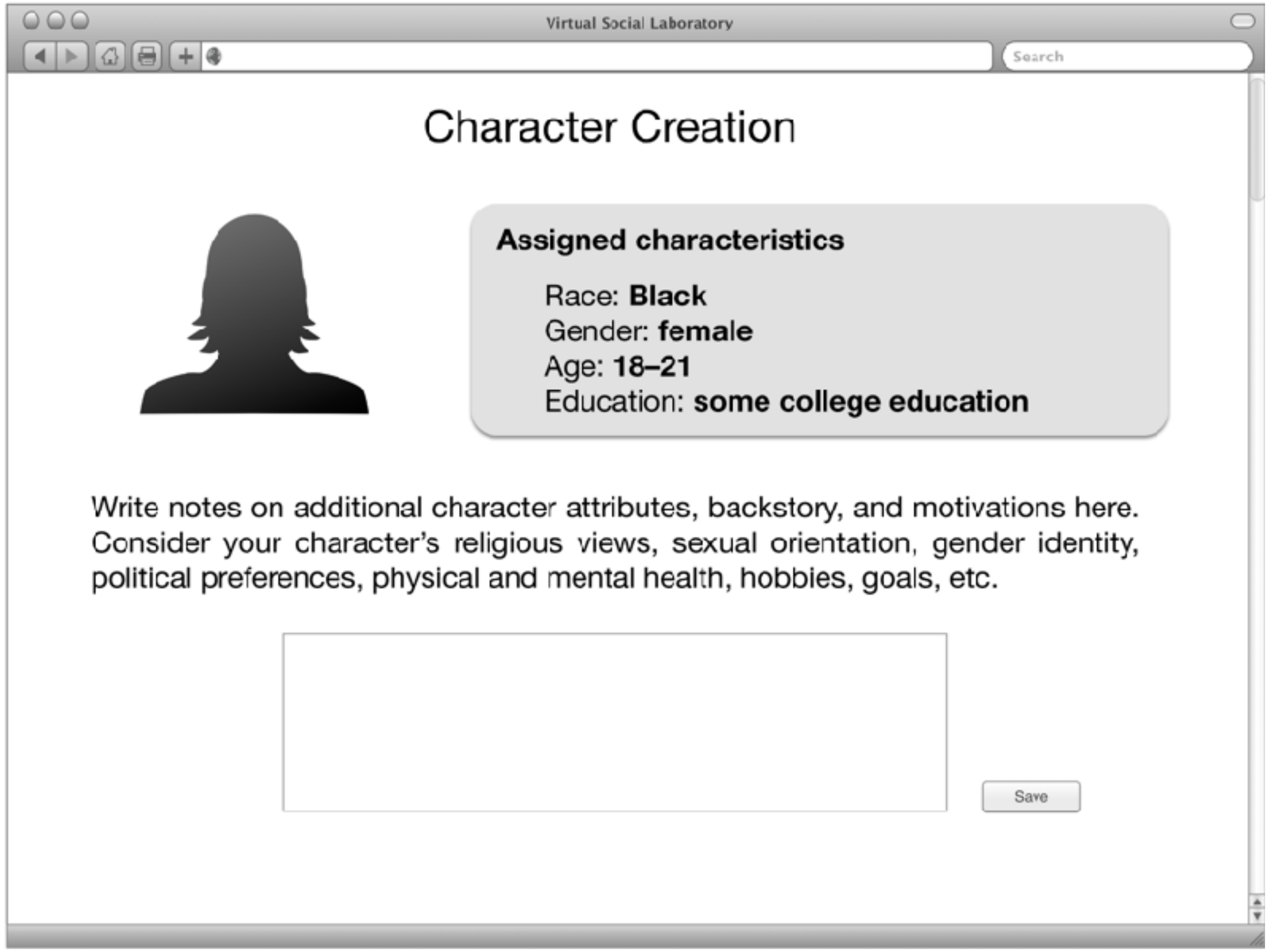
Proposal: A Virtual Social Laboratory

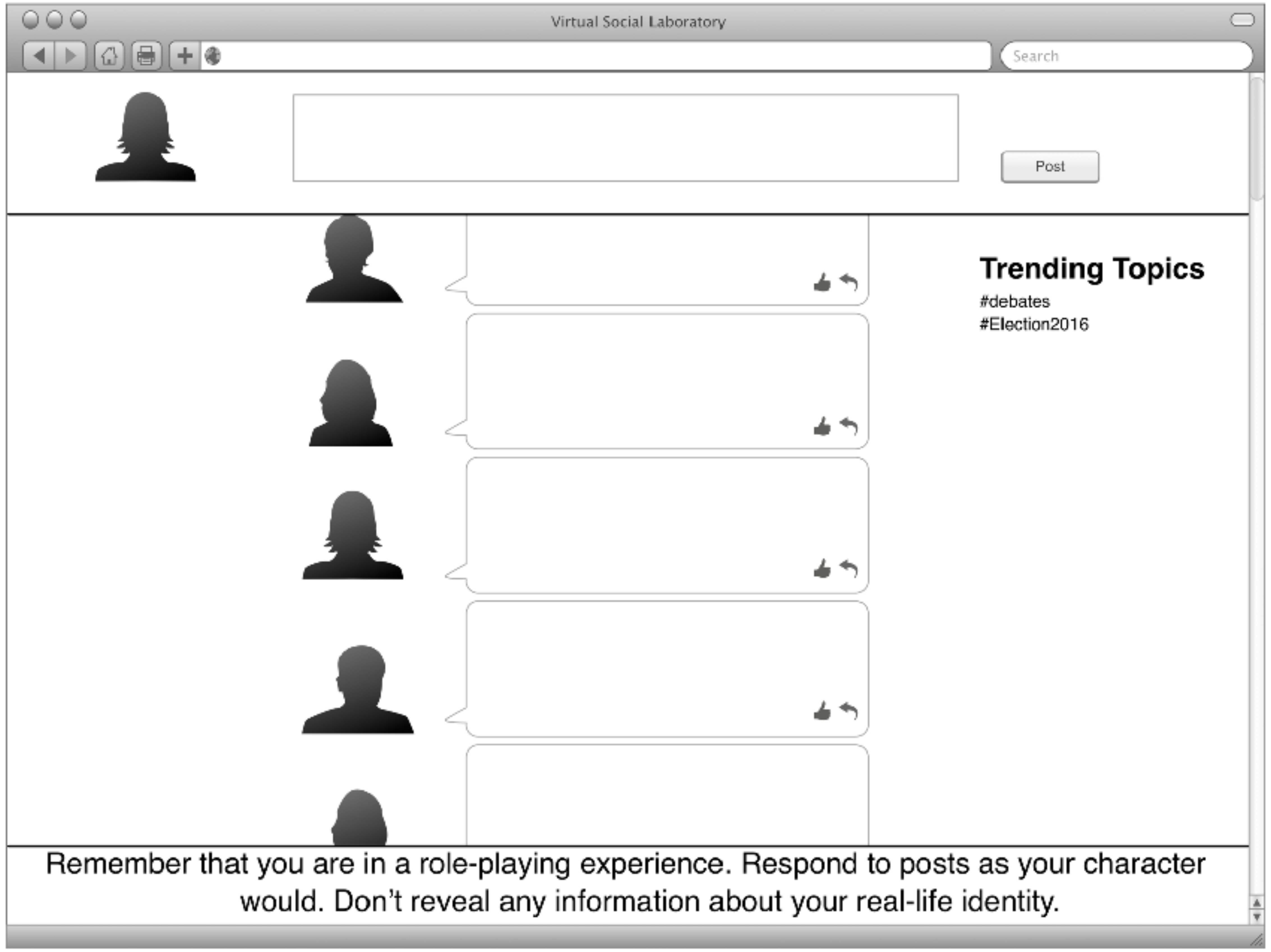
- Online social network with all users role-playing fabricated personas
- Safe environment to experiment with cybersafety technology
  - Role-playing to mitigate psychological damage of cyberviolence; protects personal identity from hybrid offline-online attacks

Proposal: A Virtual Social Laboratory

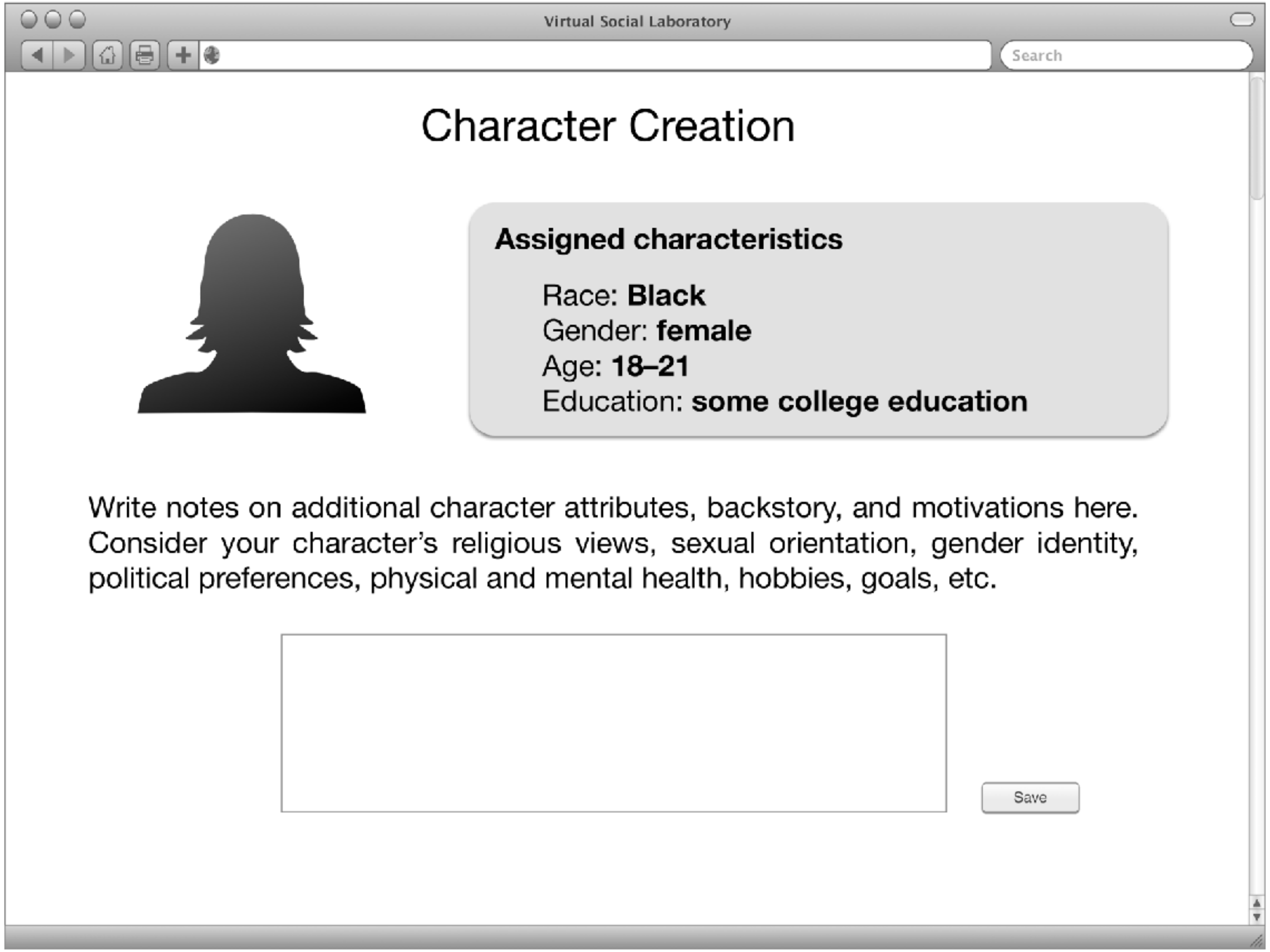
- Online social network with all users role-playing fabricated personas
- Safe environment to experiment with cybersafety technology
  - Role-playing to mitigate psychological damage of cyberviolence; protects personal identity from hybrid offline-online attacks
  - Gamified reward system to incentivize realistic play

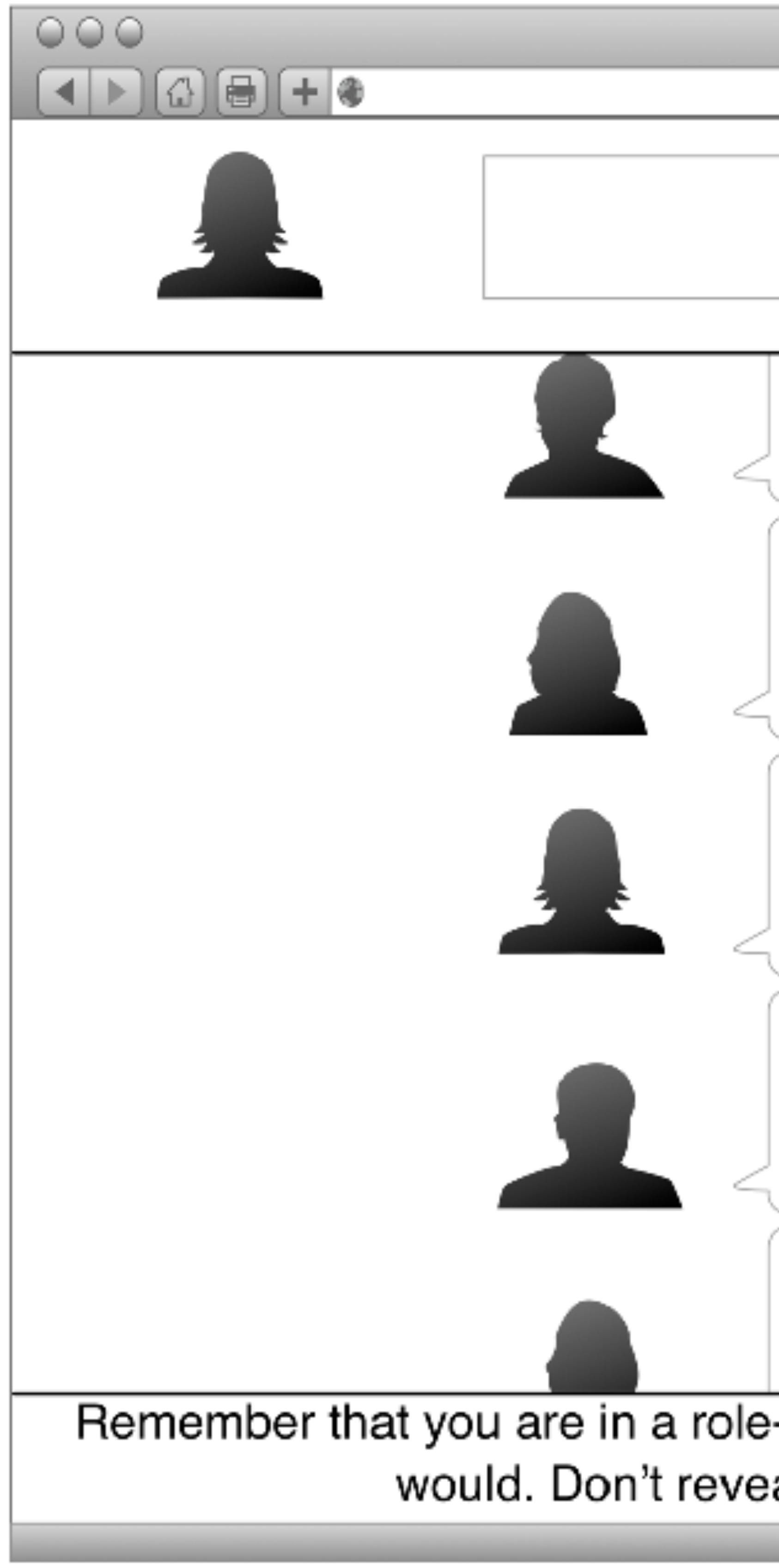
Proposal: A Virtual Social Laboratory



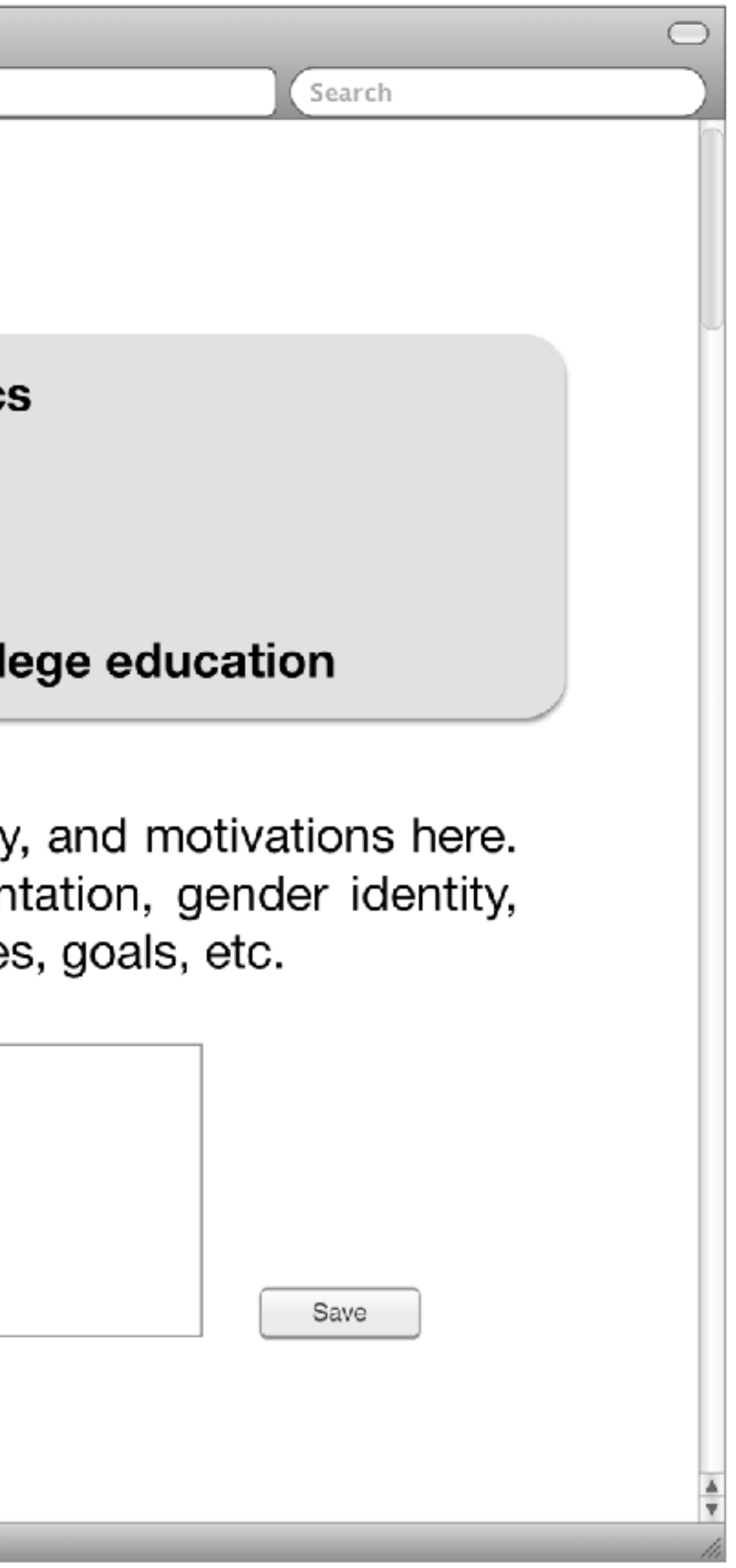


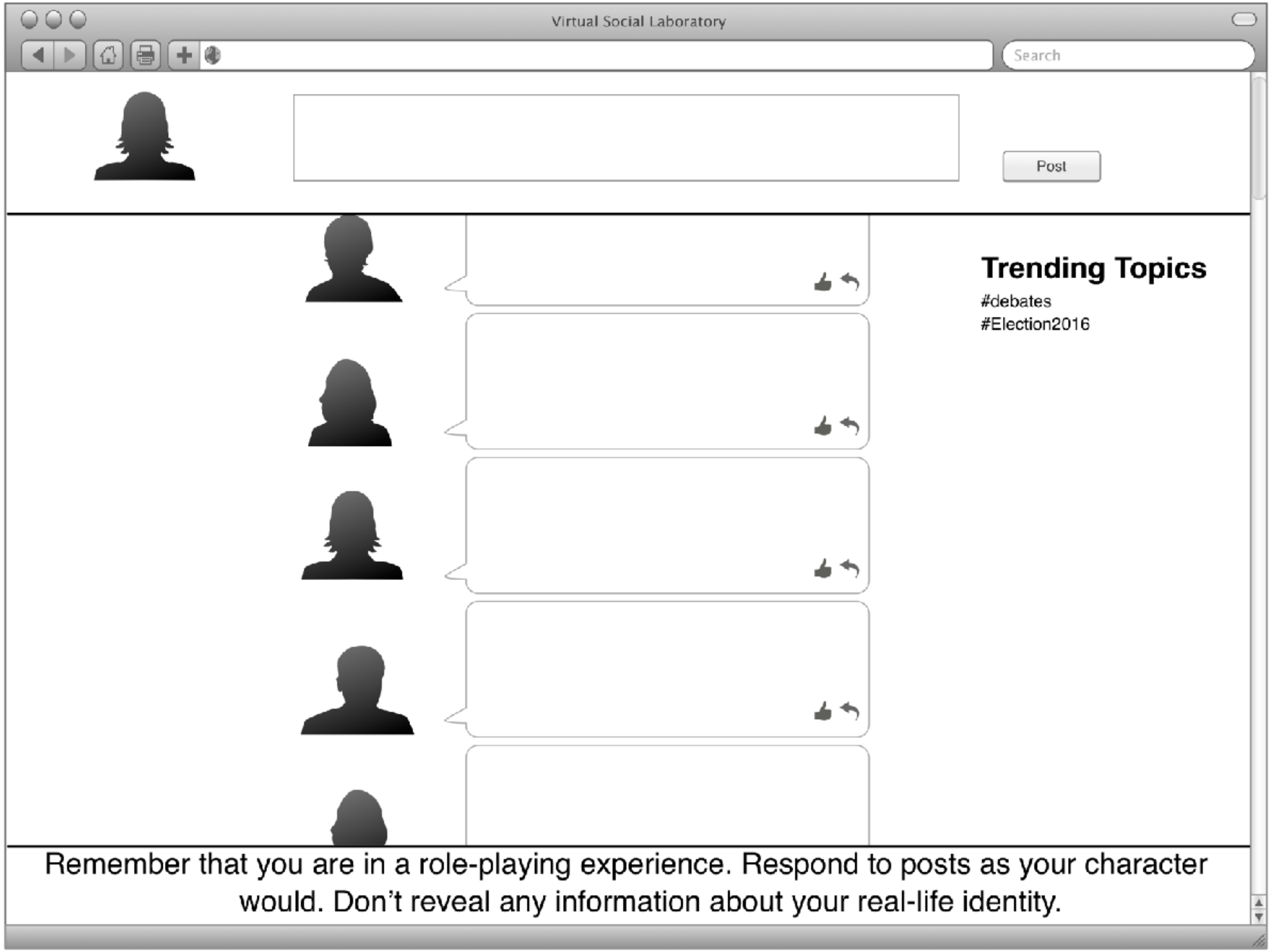
Proposal: A Virtual Social Laboratory





Proposal: A Virtual Social Laboratory





Planned Features for Virtual Social Lab

Planned Features for Virtual Social Lab

- Peer-reviewed realistic role playing

Planned Features for Virtual Social Lab

- Peer-reviewed realistic role playing
    - Large scale emulates real-world social dynamics

Planned Features for Virtual Social Lab

- Peer-reviewed realistic role playing
  - Large scale emulates real-world social dynamics
  - Role-playing emulates nuances of personal context

Planned Features for Virtual Social Lab

- Peer-reviewed realistic role playing
  - Large scale emulates real-world social dynamics
  - Role-playing emulates nuances of personal context
- Intervention experiments

Planned Features for Virtual Social Lab

- Peer-reviewed realistic role playing
    - Large scale emulates real-world social dynamics
    - Role-playing emulates nuances of personal context
- Intervention experiments
- Data collection

Planned Features for Virtual Social Lab

- Peer-reviewed realistic role playing
  - Large scale emulates real-world social dynamics
  - Role-playing emulates nuances of personal context
- Intervention experiments
- Data collection
- Measurement of sociological theories on cyberviolence

Automated Interventions

- Technology for cybersafety is aimed toward impact on social health
- Need serious thought to understand ethics and strategies for deployment and evaluation
- Proposed idea: virtual social laboratory based on role-playing

Summary & Closing Thoughts

- Challenges for machine learning approaches to detection
- New method based on weak supervision
- How to we ethically measure effectiveness before deployment?