Carnegie Mellon

MetaPaper: Lessons from 1K rejected research papers

Christos Faloutsos

CMU

https://www.cs.cmu.edu/~christos/MetaPaper

Carnegie Mellon

Disclaimer – meta-advice

Mainly, for ML / Data Science / DB venues

For other venues:

Some of the suggestions may still hold, but:

👉 Team up with seasoned authors from **those** venues

Self-introduction

Qualifications:

https://www.cs.cmu.edu/~christos/

- Over 30+ years of research-paper submissions from CMU (and UofT, UMD, AT&T, IBM, MSR, etc)
- 3-digit acceptances
- 4-digit (~1K) rejections
- …

Self-introduction

Qualifications:

https://www.cs.cmu.edu/~christos/

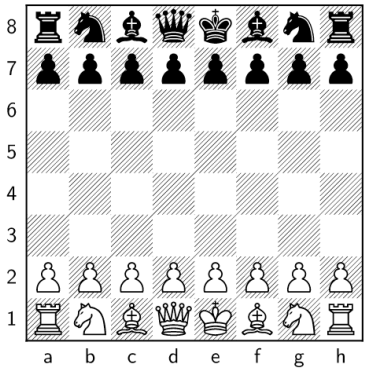
- Over 30+ years of research-paper submissions from CMU (and UofT, UMD, AT&T, IBM, MSR, etc)
- 3-digit acceptances
- 4-digit (~1K) rejections
- … and counting (sigh!)

On the bright side:

You may learn much more from a game you lose than from a game you win.

wikipedia.org

José Raúl Capablanca
World chess champion 1921-27
('the Cuban chess machine')



On the bright side:

You may learn much more from a game you lose than from a game you win.

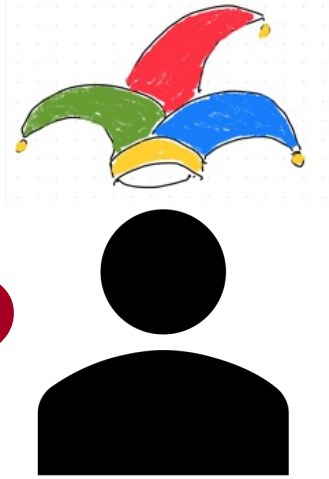
wikipedia.org

Only a fool learns from his own mistakes. The wise man learns from the mistakes of others.

(attr. to Bismarck ('Iron Chancellor') - wikiquote.org)

germanculture.com.ua

(fool?)

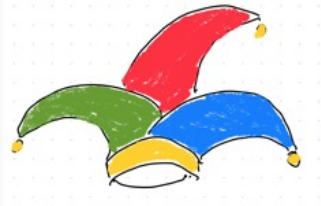


Only a fool learns from his own mistakes. The wise man learns from the mistakes of others.

(attr. to Bismarck (‘Iron Chancellor’) - wikiquote.org)

germanculture.com.ua

(ouch!)



Only a fool learns from his own mistakes. The wise man learns from the mistakes of others.

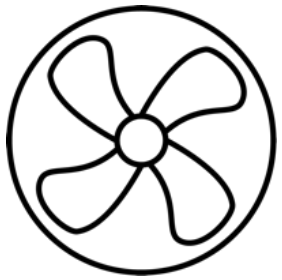
(attr. to Bismarck (‘Iron Chancellor’) - wikiquote.org)

germanculture.com.ua

Carnegie Mellon

Outline

- Top 3 lessons: ‘F.A.N.’
  - F
  - A
  - N

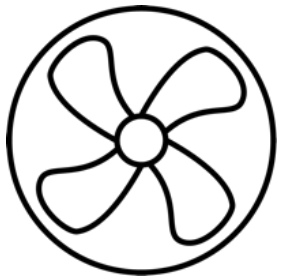


- A step back: ‘think like a reviewer’
- More battle scars; and ~10 remedies
- Conclusions

Carnegie Mellon

Outline

- Top 3 lessons: 'F.A.N.'
  - F
  - A
  - N



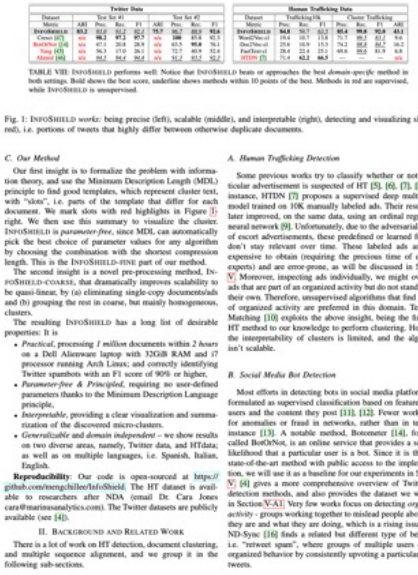
- A step back: 'think like a reviewer'
- More battle scars; and ~10 remedies
- Conclusions

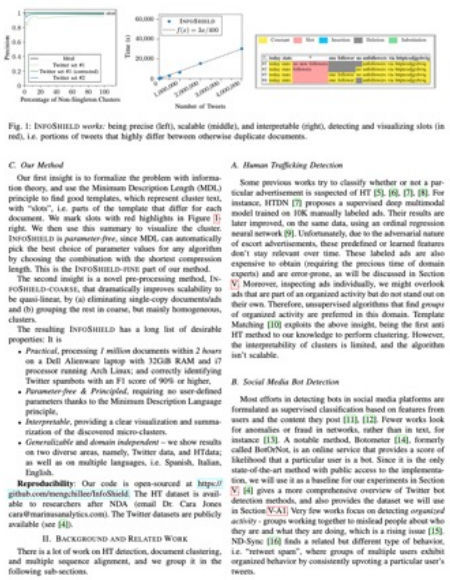
Carnegie Mellon

‘F’ – which one?

- Pick one to review
- Q1) Which one?
- Q2) Why?

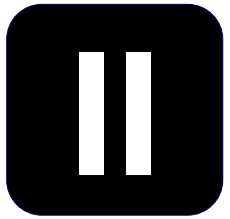
(a) (b) (c)





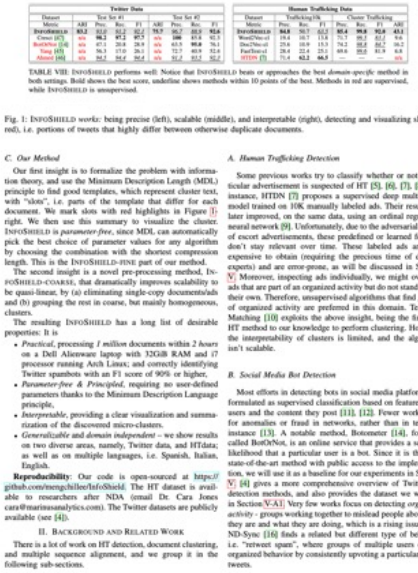
Carnegie Mellon

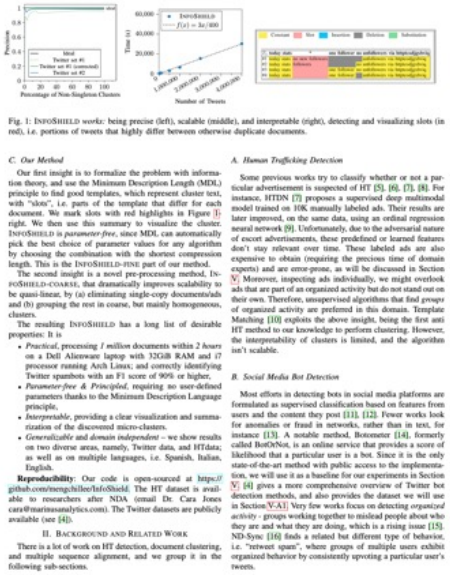
‘F’ – which one?



- Pick one to review
- Q1) Which one?
- Q2) Why?

(a) (b) (c)





Carnegie Mellon

‘F’ – which one?

- Pick one to review
- Q1) Which one? -> **‘c’**
- Q2) Why?

(a) (b) (c)

Carnegie Mellon

‘F’ – which one?

- Pick one to review
- Q1) Which one? -> ‘c’
- Q2) Why?
  - A2.1: ‘an image is worth a thousand words’
  - A2.2: the author respects reviewer’s time

(c)

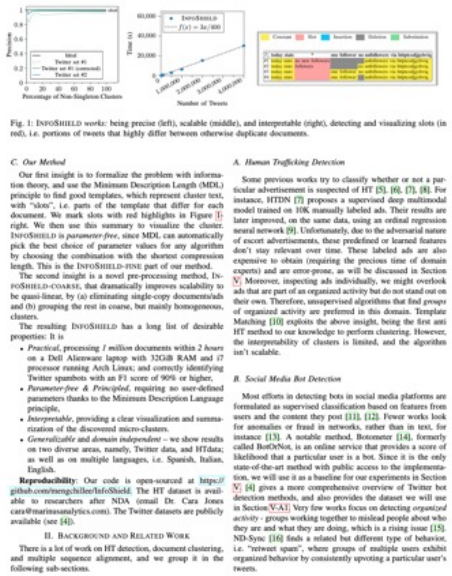


Fig. 1: INFOSHIELD *works:* being precise (left), scalable (middle), and interpretable (right), detecting and visualizing slots (in red), i.e. portions of tweets that highly differ between otherwise duplicate documents.

C. Our Method

Our first insight is to formalize the problem with information theory, and use the Minimum Description Length (MDL) principle to find good templates, which represent cluster text, with “slots”, i.e. parts of the template that differ for each document. We mark slots with red highlights in Figure 1: right. We then use this summary to visualize the cluster. INFOSHIELD is *parameter-free*, since MDL can automatically pick the best choice of parameter values for any algorithm by choosing the combination with the shortest compression length. This is the INFOSHIELD-FINE part of our method.

The second insight is a novel pre-processing method, INFOSHIELD-COARSE, that dramatically improves scalability to be quasi-linear, by (a) eliminating single-copy documents/ads and (b) grouping the rest in coarse, but mainly homogeneous, clusters.

The resulting INFOSHIELD has a long list of desirable properties: It is

- *Practical*, processing *1 million* documents within *2 hours* on a Dell Alienware laptop with 32GiB RAM and i7 processor running Arch Linux; and correctly identifying Twitter spambots with an F1 score of 90% or higher,
- *Parameter-free & Principled*, requiring no user-defined parameters thanks to the Minimum Description Language principle,
- *Interpretable*, providing a clear visualization and summarization of the discovered micro-clusters.
- *Generalizable and domain independent* – we show results on two diverse areas, namely, Twitter data, and HTdata; as well as on multiple languages, i.e. Spanish, Italian, English.

Reproducibility: Our code is open-sourced at https://github.com/mengchillee/InfoShield. The HT dataset is available to researchers after NDA (email Dr. Cara Jones cara@marinsanalytics.com). The Twitter datasets are publicly available (see [4]).

II. BACKGROUND AND RELATED WORK

There is a lot of work on HT detection, document clustering, and multiple sequence alignment, and we group it in the following sub-sections.

A. Human Trafficking Detection

Some previous works try to classify whether or not a particular advertisement is suspected of HT [5], [6], [7], [8]. For instance, HTDN [7] proposes a supervised deep multimodal model trained on 10K manually labeled ads. Their results are later improved, on the same data, using an ordinal regression neural network [9]. Unfortunately, due to the adversarial nature of escort advertisements, these predefined or learned features don’t stay relevant over time. These labeled ads are also expensive to obtain (requiring the precious time of domain experts) and are error-prone, as will be discussed in Section V. Moreover, inspecting ads individually, we might overlook ads that are part of an organized activity but do not stand out on their own. Therefore, unsupervised algorithms that find groups of organized activity are preferred in this domain. Template Matching [10] exploits the above insight, being the first anti HT method to our knowledge to perform clustering. However, the interpretability of clusters is limited, and the algorithm isn’t scalable.

B. Social Media Bot Detection

Most efforts in detecting bots in social media platforms are formulated as supervised classification based on features from users and the content they post [11], [12]. Fewer works look for anomalies or fraud in networks, rather than in text, for instance [13]. A notable method, Botometer [14], formerly called BotOrNot, is an online service that provides a score of likelihood that a particular user is a bot. Since it is the only state-of-the-art method with public access to the implementation, we will use it as a baseline for our experiments in Section V. [4] gives a more comprehensive overview of Twitter bot detection methods, and also provides the dataset we will use in Section V-A1. Very few works focus on detecting *organized activity* - groups working together to mislead people about who they are and what they are doing, which is a rising issue [15]. ND-Sync [16] finds a related but different type of behavior, i.e. “retweet spam”, where groups of multiple users exhibit organized behavior by consistently upvoting a particular user’s tweets.

Carnegie Mellon

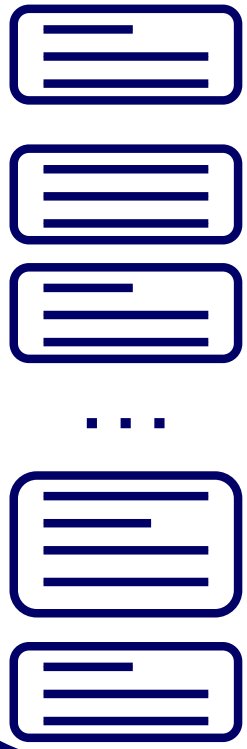
Outline

- Top 3 lessons: ‘F.A.N.’
  - Figure (‘crown jewel’ figure)
  - A
  - N

- A step back: ‘think like a reviewer’
- More battle scars; and ~10 remedies
- Conclusions

'F': What makes a good 'crown jewel' figure?

- [Consider a paper that tries to find near-duplicate escort ads -> human trafficking]



Carnegie Mellon

(paper)

Meng-Chieh Lee, Catalina Vajiac, Aayushi Kulshrestha, Sacha Levy, Namyong Park, Cara Jones, Reihaneh Rabbany, and Christos Faloutsos

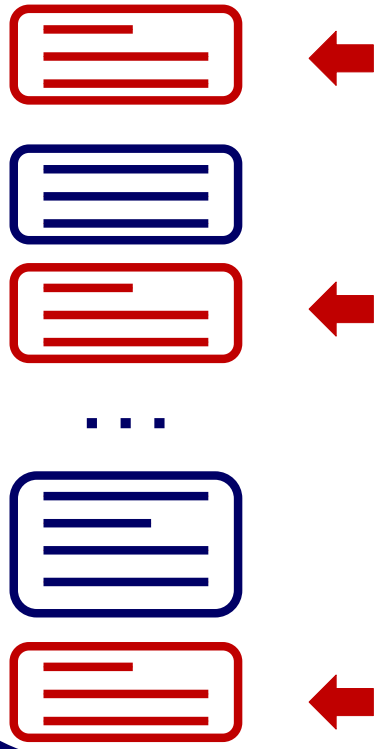
INFOSHIELD: Generalizable Information-Theoretic Human-Trafficking Detection

ICDE 2021, Chania, Greece, April 2021.

Carnegie Mellon

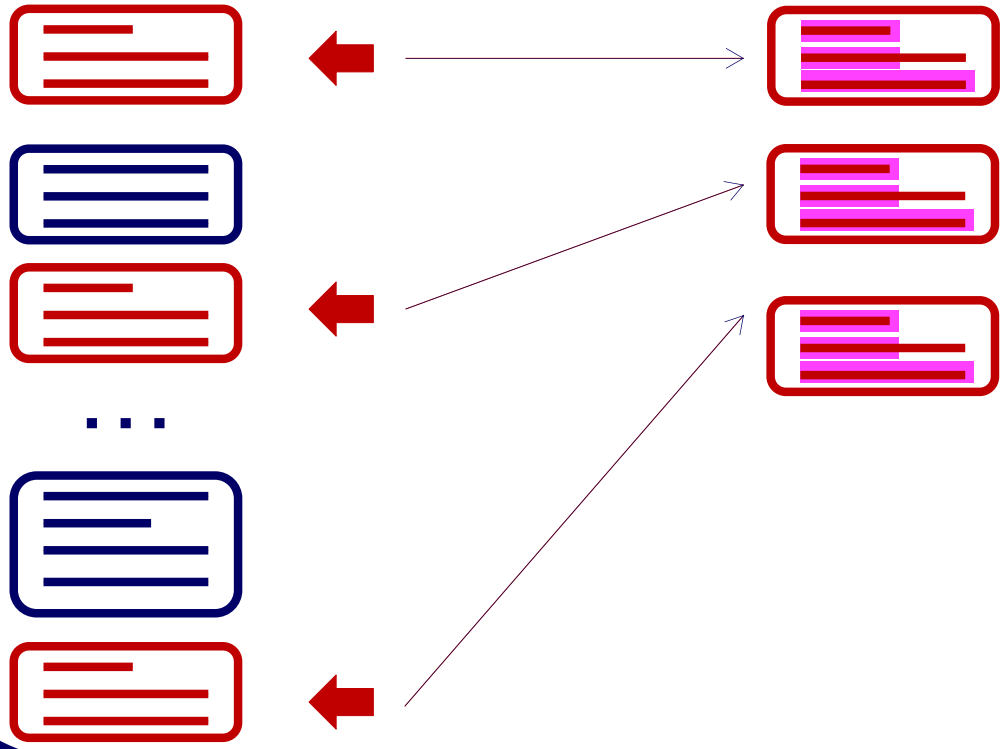
'F': What makes a good 'crown jewel' figure?

- [Consider a paper that tries to find near-duplicate escort ads -> human trafficking]



'F': What makes a good 'crown jewel' figure?

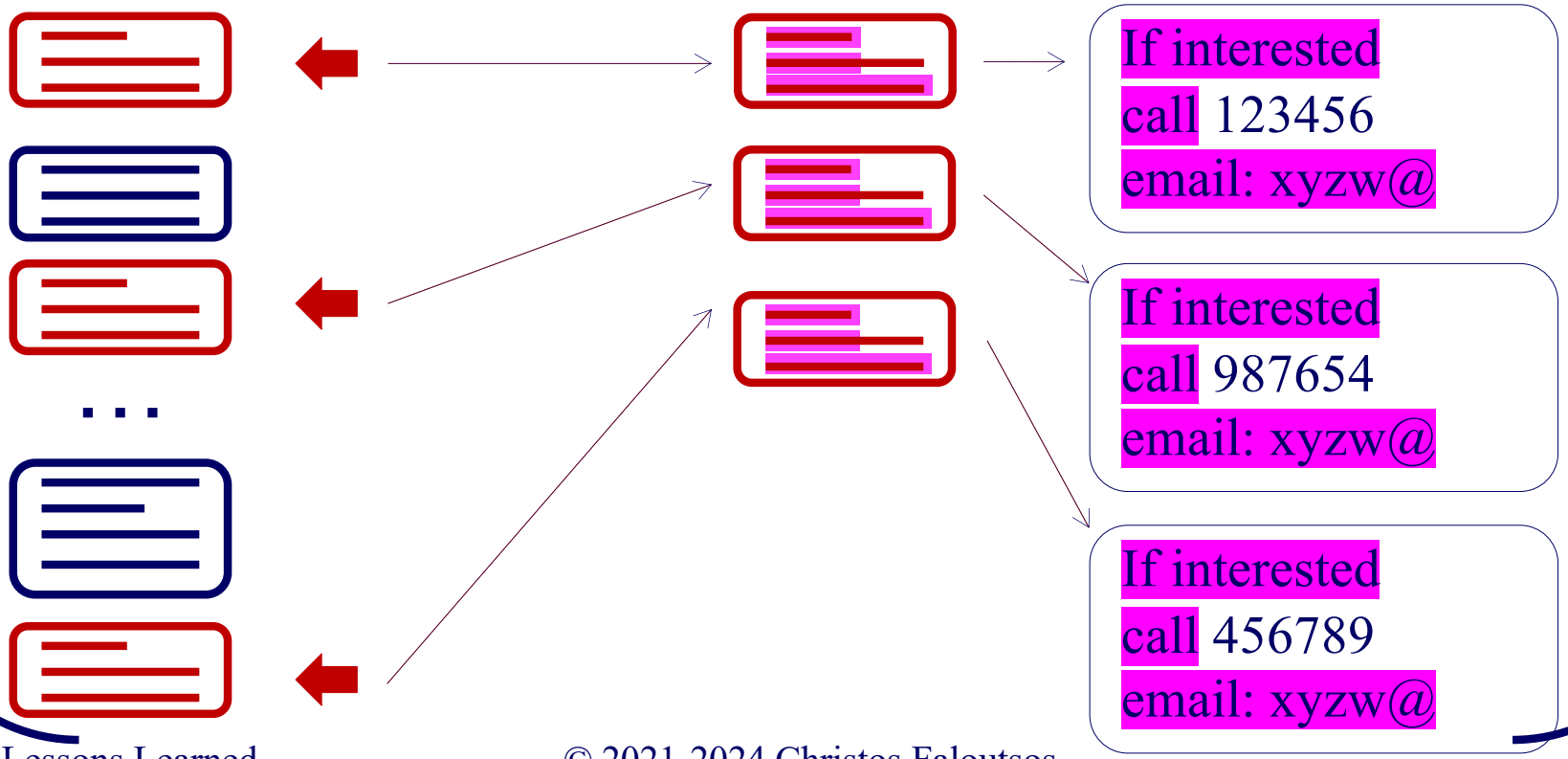
- [Consider a paper that tries to find near-duplicate escort ads -> human trafficking]



Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

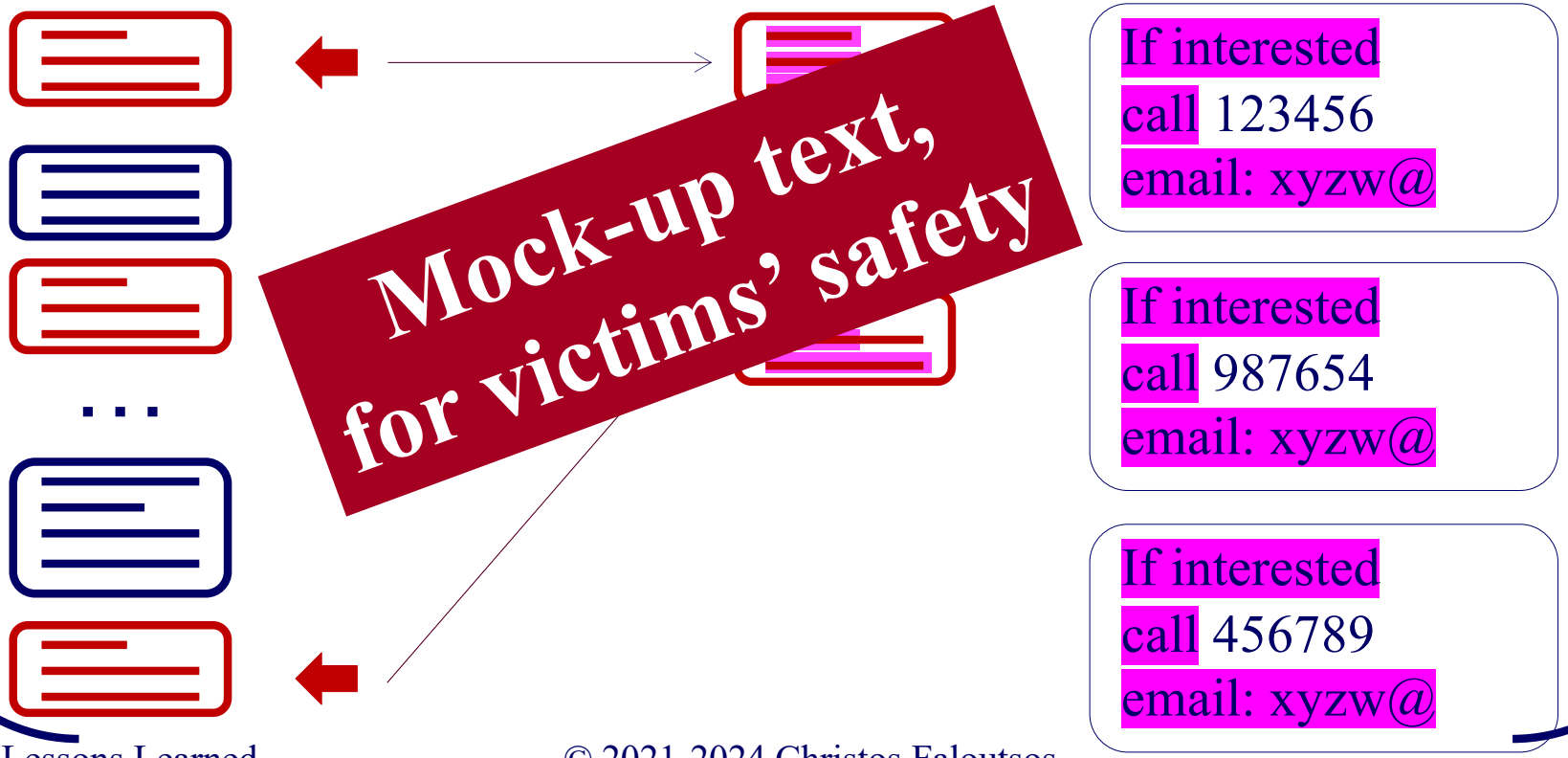
- [Consider a paper that tries to find near-duplicate escort ads -> human trafficking]



Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

- [Consider a paper that tries to find near-duplicate escort ads -> human trafficking]

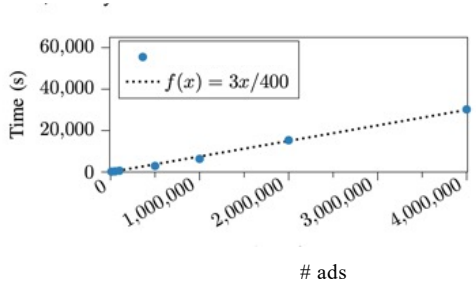


Carnegie Mellon

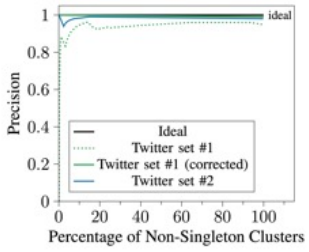
'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

A)



B)



C)

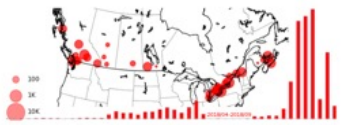
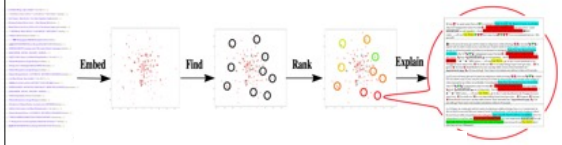
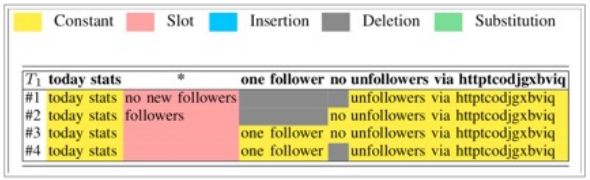


Fig. 1: *Pervasiveness of Issue:* Distribution of escort ads in

D)



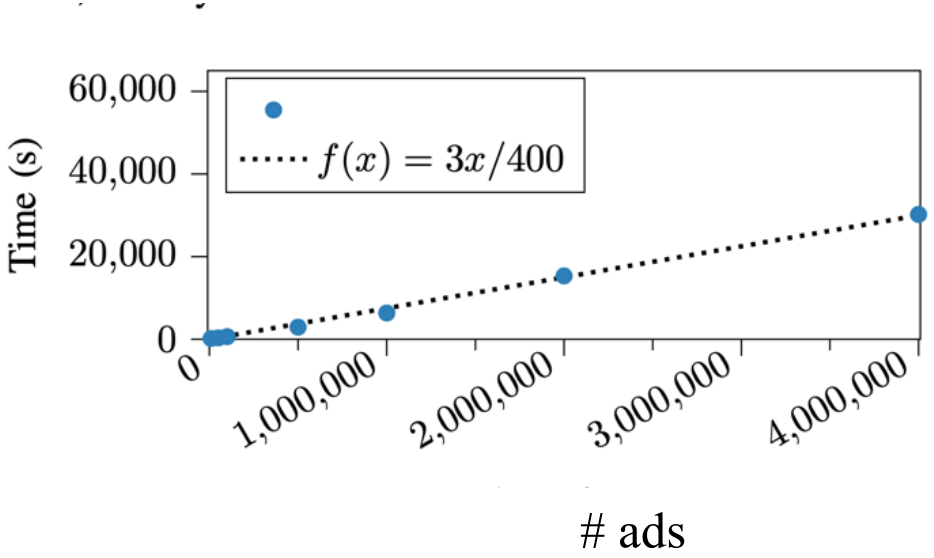
E)



'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(A)

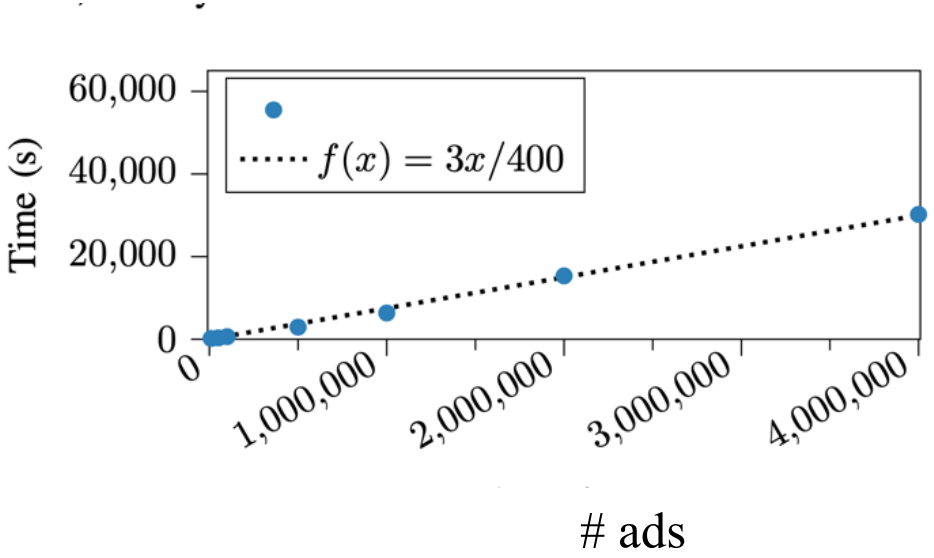


Message??

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(A)

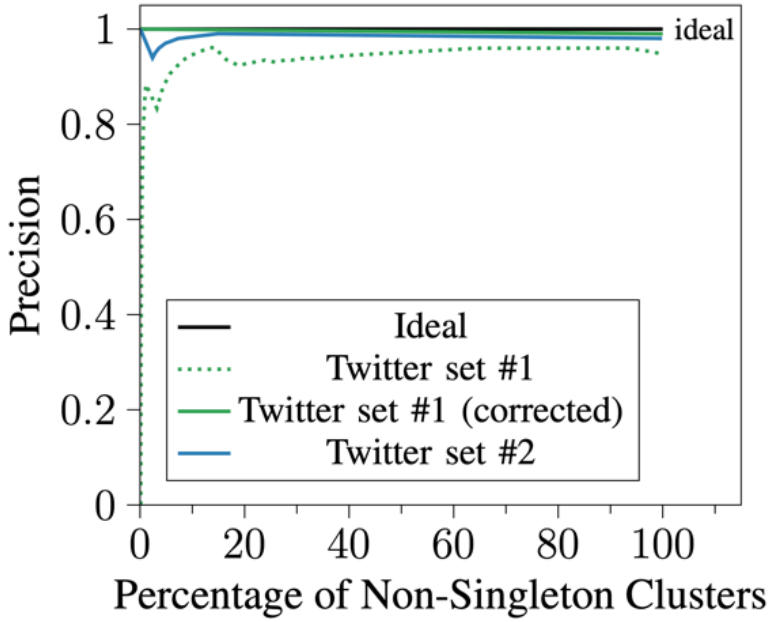


scalable

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(B)



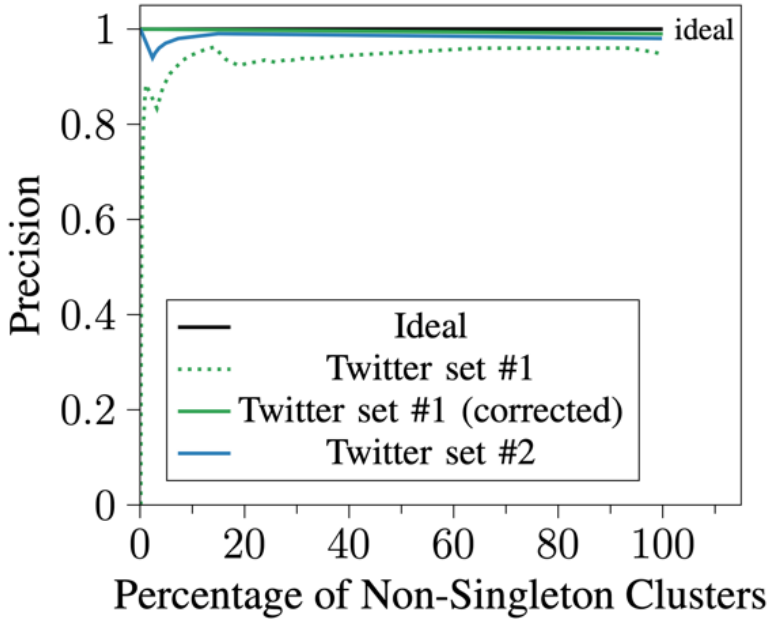
Message?

Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(B)

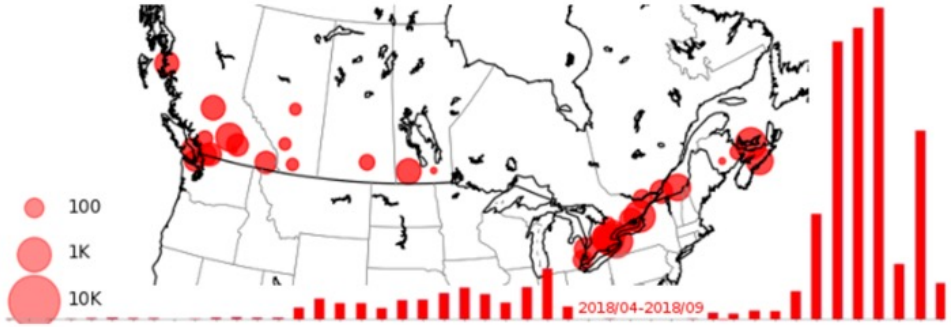


accurate

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(C)



Message?

Fig. 1: *Pervasiveness of Issue:* Distribution of escort ads in

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(C)

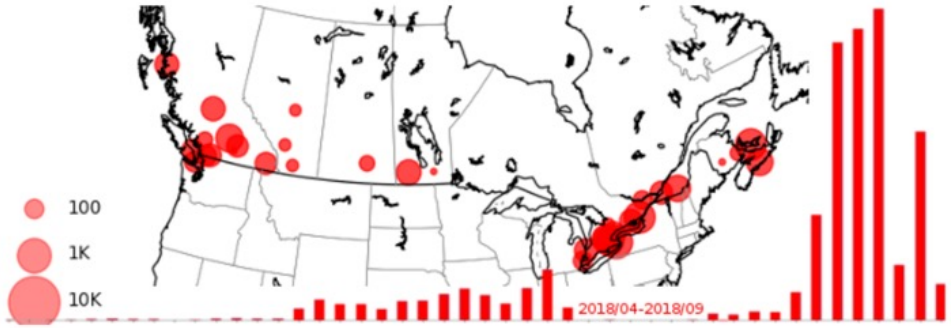


Fig. 1: *Pervasiveness of Issue:* Distribution of escort ads in

Important problem

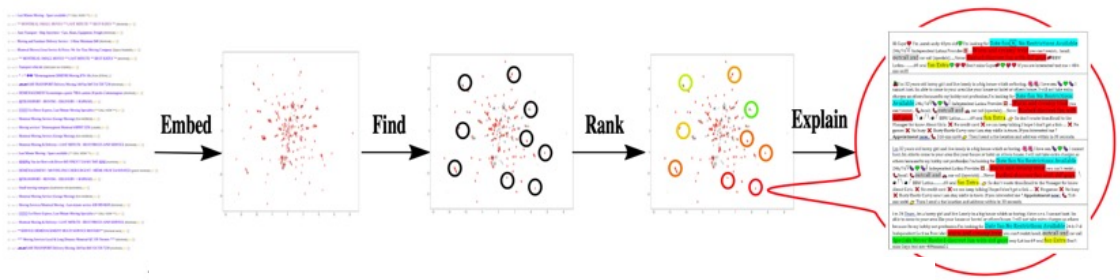
Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

Message?

(D)



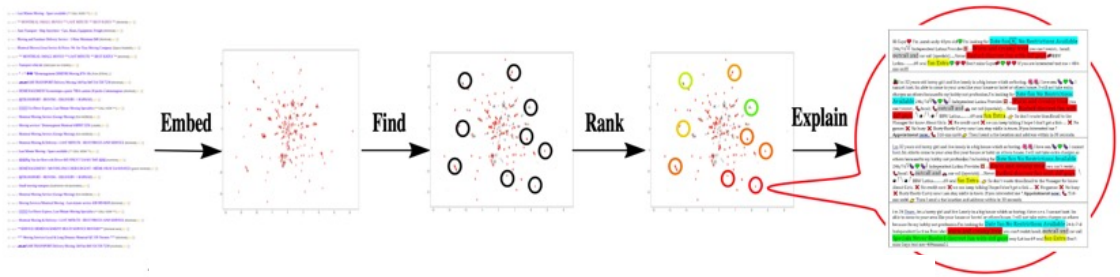
System Overview

Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(D)



System Overview

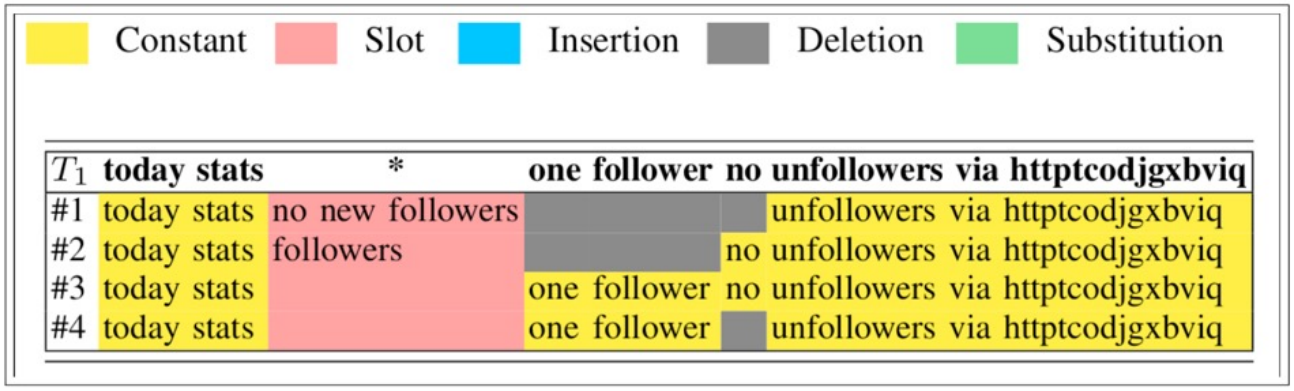
Elaborate method

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(E)

Message?



T_1	today stats	*	one follower	no unfollowers via httptcodjgxbviq
#1	today stats	no new followers		unfollowers via httptcodjgxbviq
#2	today stats	followers		no unfollowers via httptcodjgxbviq
#3	today stats		one follower	no unfollowers via httptcodjgxbviq
#4	today stats		one follower	unfollowers via httptcodjgxbviq

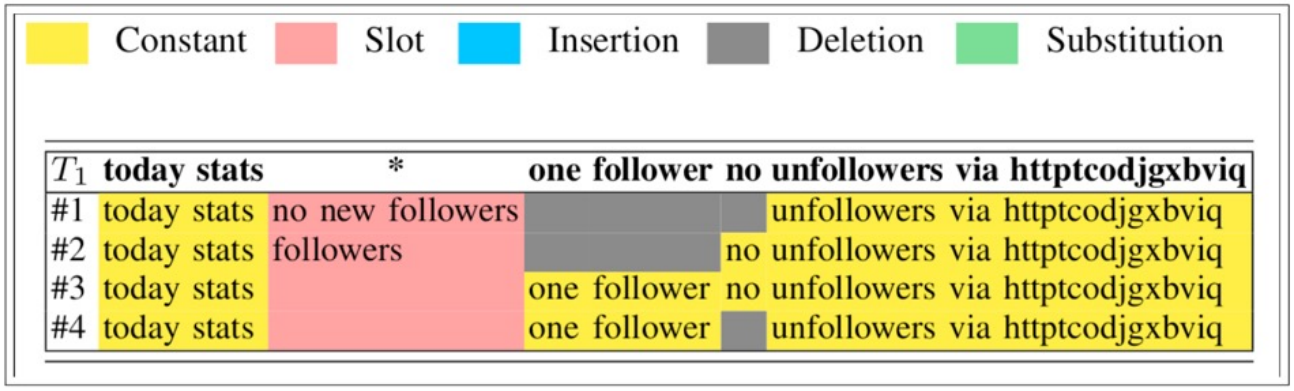
Result, on real 'tweets'

'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

(E)

It works.



T_1	today stats	*	one follower	no unfollowers via httptcodjgxbviq
#1	today stats	no new followers		unfollowers via httptcodjgxbviq
#2	today stats	followers		no unfollowers via httptcodjgxbviq
#3	today stats		one follower	no unfollowers via httptcodjgxbviq
#4	today stats		one follower	unfollowers via httptcodjgxbviq

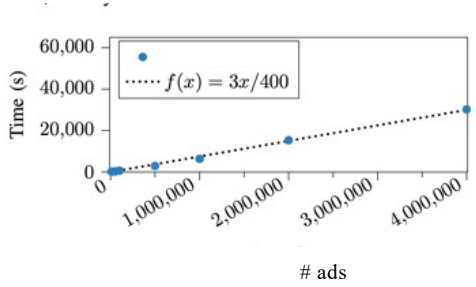
Result, on real 'tweets'

Carnegie Mellon

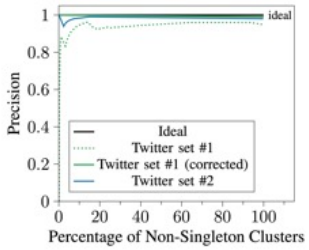
'F': What makes a good 'crown jewel' figure?

- Which one a reviewer will find most impressive?

A) scalable



B) accurate



C) important

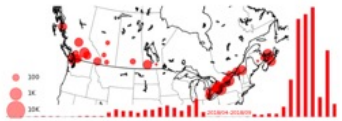
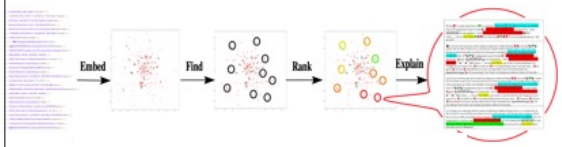
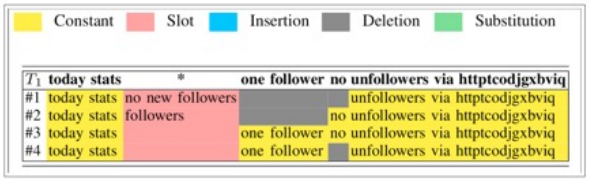


Fig. 1: *Pervasiveness of Issue:* Distribution of escort ads in

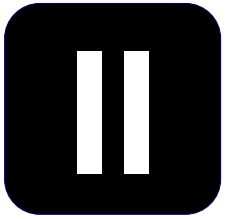
D) elaborate



E) It works

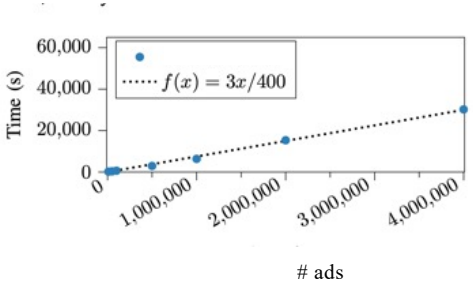


'F': What makes a good 'crown jewel' figure?

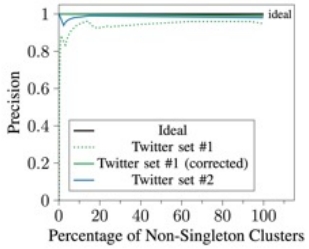


- Which one a reviewer will find most impressive?

A) scalable



B) accurate



C) important

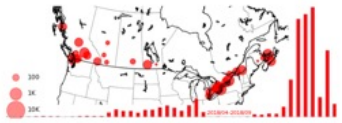
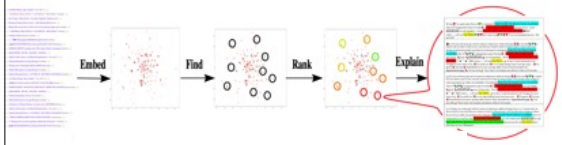
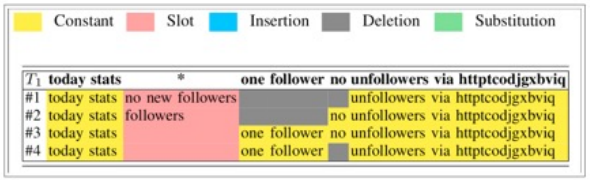


Fig. 1: *Pervasiveness of Issue:* Distribution of escort ads in

D) elaborate



E) It works

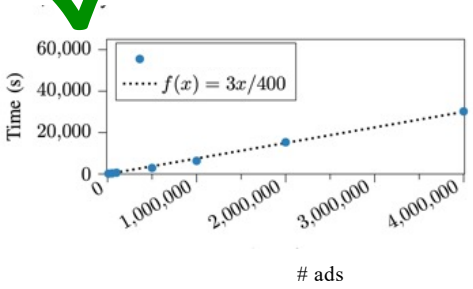


Carnegie Mellon

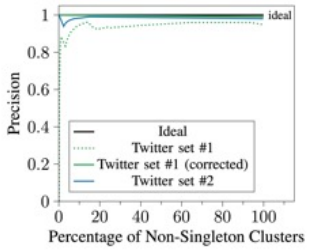
'F': What makes a good 'crown jewel' figure?

- Q: Which one to choose?

A) scalable



B) accurate



C) important

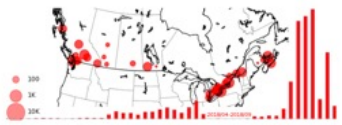
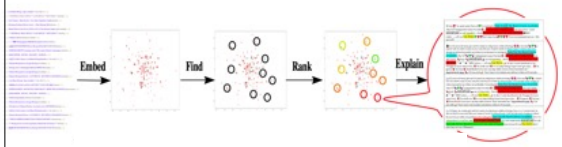
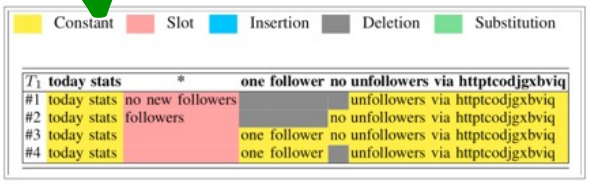


Fig. 1: *Pervasiveness of Issue:* Distribution of escort ads in

D) elaborate



E) It works

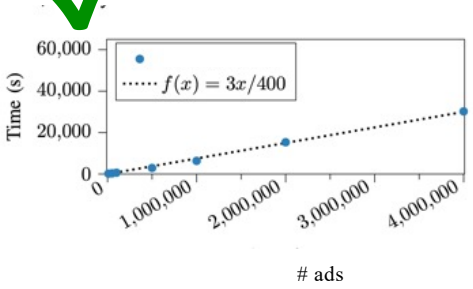


Carnegie Mellon

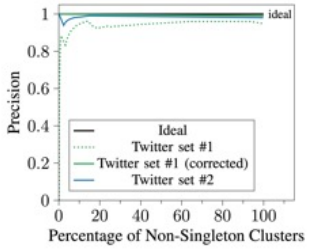
'F': What makes a good 'crown jewel' figure?

- A: all 3 - it's OK to have >1 figures

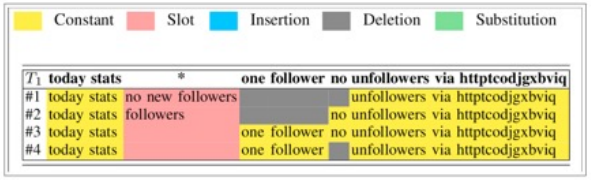
✓ A) scalable



✓ B) accurate



✓ E) It works



Carnegie Mellon

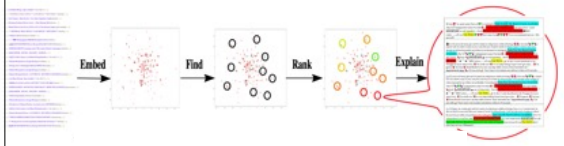
'F': What makes a good 'crown jewel' figure?

- Emphasize '***what***' we can achieve – not '***how***'

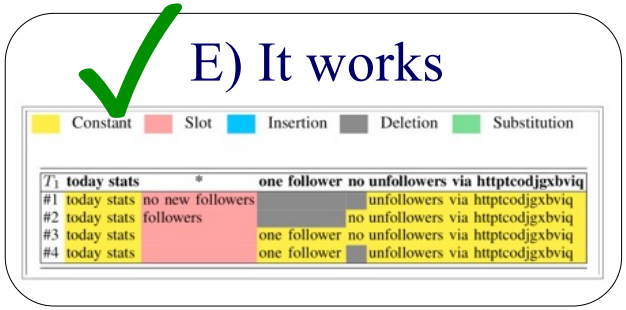
'How'

'What'

D) elaborate



E) It works



Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

- Emphasize '***what***' we can achieve – not '***how***'

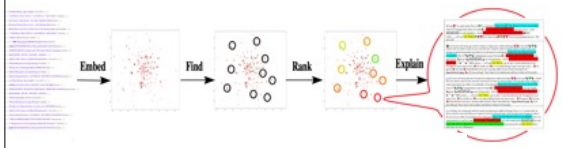
'How'

Embed -> cluster -> rank"

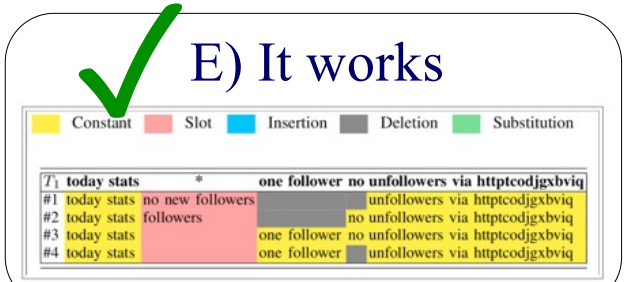
*"*anybody* could do that – what's so novel?"*

'What'

D) elaborate



E) It works



Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

- Emphasize '***what***' we can achieve – not '***how***'

Embed

"*anybo

wha

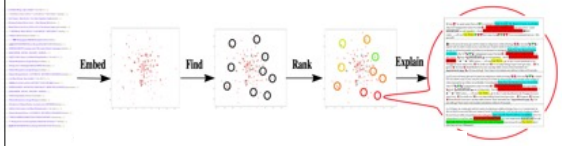
-> rank"

lo that –

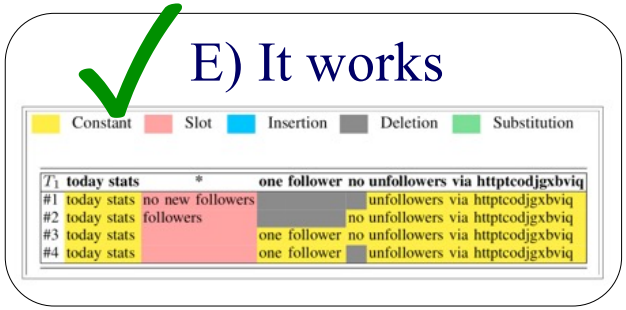
l?"

'What'

D) elaborate



E) It works



Carnegie Mellon

'F': What makes a good 'crown jewel' figure?

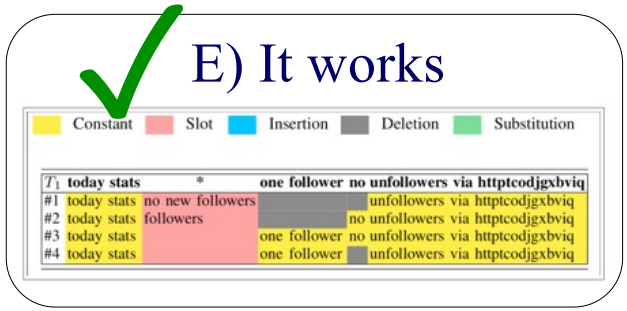
- Emphasize '***what***' we can achieve – not '***how***'

Emb rank"

"*any *that –*

v "

'What'



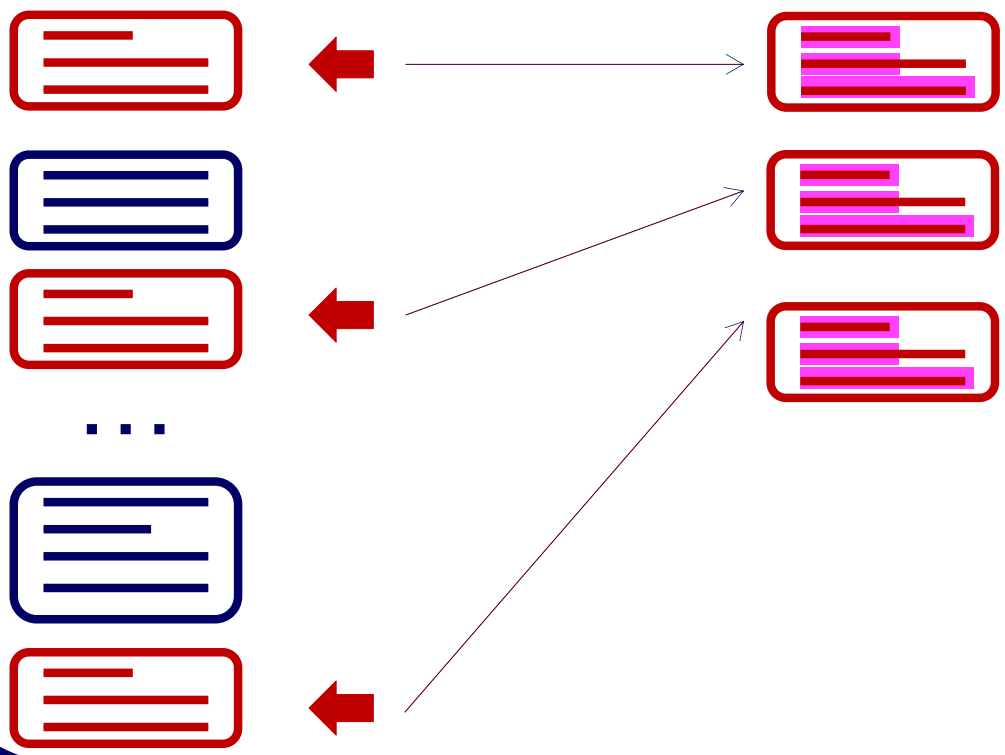
Outline

- Top 3 lessons: ‘F.A.N.’
  - Figure (‘crown jewel’ figure)
  - A
  - N

- A step back: ‘think like a reviewer’
- More battle scars; and ~10 remedies
- Conclusions

'A': Best way to start the abstract?

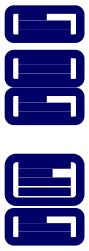
- [Consider a paper that tries to find near-duplicate escort ads -> human trafficking]



If interested
call 123456
email: xyzw@

If interested
call 987654
email: xyzw@

If interested
call 456789
email: xyzw@

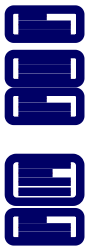


'A': Best way to start the abstract?

A)

Human trafficking is an age old problem that continues to affect 25 million people worldwide.

…..



'A': Best way to start the abstract?

A)

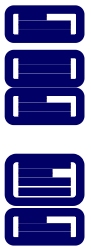
Human trafficking is an age old problem that continues to affect 25 million people worldwide.

…..

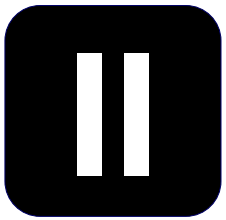
B)

Given a million escort advertisements, how can we spot near- duplicates?

…..



'A': Best way to start the abstract?



A)

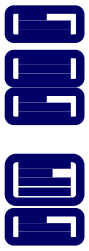
Human trafficking is an age old problem that continues to affect 25 million people worldwide.

…..

B)

Given a million escort advertisements, how can we spot near- duplicates?

…..



'A': Best way to start the abstract?

✓ A)

Human trafficking is an age old problem that continues to affect 25 million people worldwide.

…..

✓✓ B)

Given a million escort advertisements, how can we spot near- duplicates?

…..

Outline

- Top 3 lessons: 'F.A.N.'
  - Figure ('crown jewel' figure)
  - Ask (ask a rhetorical question)
  - N

- A step back: 'think like a reviewer'
- More battle scars; and ~10 remedies
- Conclusions

'A': Best way to start the abstract?

A)

Human trafficking is an age old problem that continues to affect 25 million people worldwide.

…..

B)

Given a million escort advertisements, how can we spot near- duplicates?

…..

'Rhetorical question':
= you know the answer

'A': Best way to start the abstract?

A)

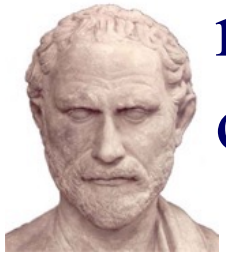
Human trafficking is an age old problem that continues to affect 25 million people worldwide.

…..

B)

Given a million escort advertisements, how can we spot near- duplicates?

…..



rhetor
orator

'Rhetorical question':
= you know the answer

'A': Best 'rhetorical question':

C)

How can we help law enforcement fight human trafficking?

…..

B)

Given a million escort advertisements, how can we spot near- duplicates?

…..

'Rhetorical question':
= you know the answer

'A': Best 'rhetorical question':

C)

How can we help law enforcement fight human trafficking?

…..

Wrong question.
('call your congressperson', 'deploy face recognition s/w', …)

B)

Given a million escort advertisements, how can we spot near- duplicates?

…..

'Rhetorical question':
= you know the answer

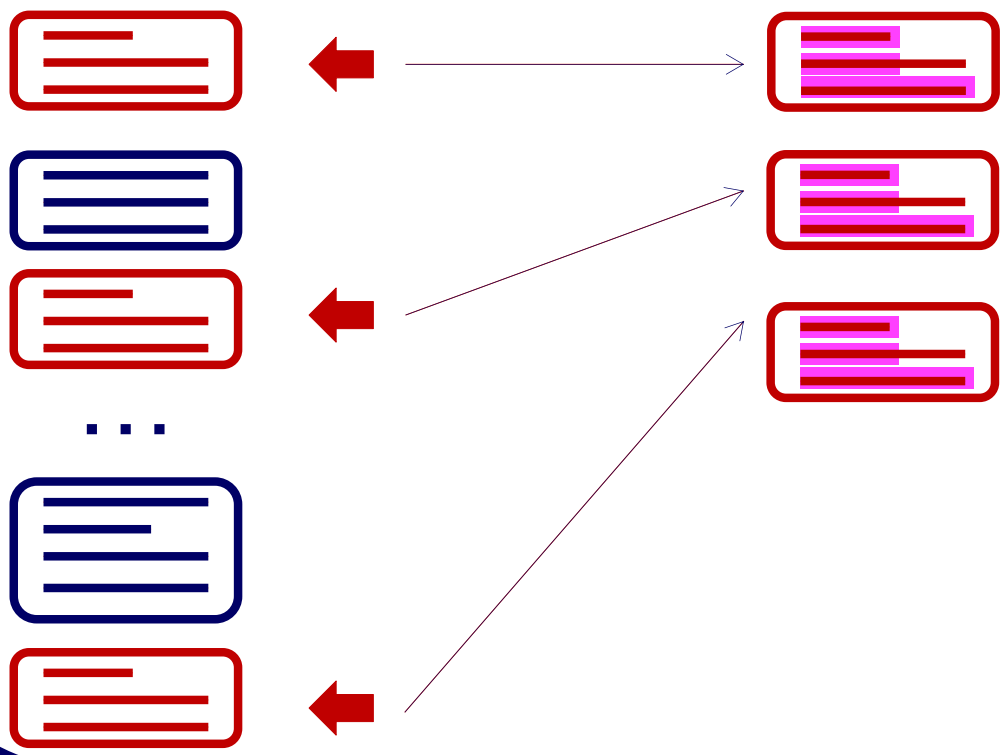
Outline

- Top 3 lessons: 'F.A.N.'
  - Figure ('crown jewel' figure)
  - Ask (ask a rhetorical question)
  - N

- A step back: 'think like a reviewer'
- More battle scars; and ~10 remedies
- Conclusions

Reminder: our sample problem

- [Consider a paper that tries to find near-duplicate escort ads -> human trafficking]

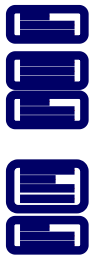


If interested
call 123456
email: xyzw@

If interested
call 987654
email: xyzw@

If interested
call 456789
email: xyzw@

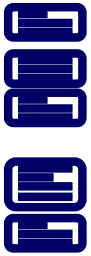
Carnegie Mellon



'N': Which title is best?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon



'N': Which title is best?

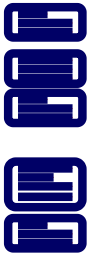
A. On human trafficking detection

B. Embedding and clustering for human trafficking detection

C. Fast and accurate human trafficking detection

D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon



'N': Which title is best?

~ A. On human trafficking detection

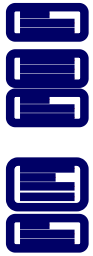
Too general

B. Embedding and clustering for human trafficking detection

C. Fast and accurate human trafficking detection

D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon



'N': Which title is best?

A. On human trafficking detection

B. Embedding and clustering for human trafficking detection

C. Fast and accurate human trafficking detection

D. TrafficSpot: Fast and accurate human trafficking detection

'N': Which title is best?

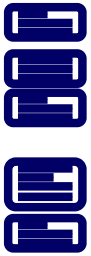
NO-NO
'What' – not 'How'

A. On human trafficking dete

B. Embedding and clustering for human trafficking detection

C. Fast and accurate human trafficking detection

D. TrafficSpot: Fast and accurate human trafficking detection



'N': Which title is best?

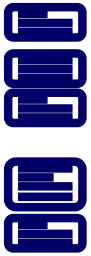
A. On human trafficking detection

B. Embedding and clustering for human trafficking detection

C. Fast and accurate human trafficking detection

D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon



'N': Which title is best?

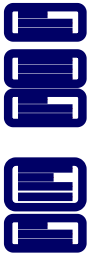
A. On human trafficking detection

B. Embedding and clustering for human trafficking detection

✓ C. Fast and accurate human trafficking detection

D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon



'N': Which title is best?

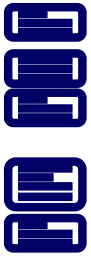
A. On human trafficking detection

B. Embedding and clustering for human trafficking detection

C. Fast and accurate human trafficking detection

D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon



'N': Which title is best?

A. On human trafficking detection

B. Embedding and clustering for human trafficking detection

C. Fast and accurate human trafficking detection

✓✓D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon

'N': Which title is best?

Q: Benefits of a name?

✓✓D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon

'N': Which title is best?

Q: Benefits of a name?

- Easy reference
- .

√√D. TrafficSpot: Fast and accurate human trafficking detection

Carnegie Mellon

'N': Which title is best?

Q: Benefits of a name?

- Easy reference
- 'novelty'

✓✓D. TrafficSpot: Fast and accurate human trafficking detection

Outline

- Top 3 lessons: 'F.A.N.'
  - Figure ('crown jewel' figure)
  - Ask (ask a rhetorical question)
  - Name (give a name to your method/system)
- A step back: 'think like a reviewer'
- More battle scars; and ~10 remedies
- Conclusions

Carnegie Mellon

'N': How to pick good names?

a) CAE (-> Cluster And Embed)

b) CUBE (-> **C**l**U**ster and em**BE**d)

c) Spot

d) TrafficSpot

e) InfoShield

'N': How to pick good names?

- a) CAE (-> Cluster And Embed)
- b) CUBE (-> **C**l**U**ster and em**BE**d)
- c) Spot
- d) TrafficSpot
- e) InfoShield

- Meaningless
- 'how'

'N': How to pick good names?

- a) CAE (-> Cluster And Embed)
- b) CUBE (-> **C**l**U**ster and em**BE**d)
- c) Spot
- d) TrafficSpot
- e) InfoShield

'N': How to pick good names?

- a) CAE (-> Cluster And Embed)
- b) CUBE (-> **C**l**U**ster and em**BE**d)
- c) Spot
- d) TrafficSpot
- e) InfoShield

- English word
- irrelevant
- 'how'

'N': How to pick good names?

- a) CAE (-> Cluster And Embed)
- b) CUBE (-> **C**l**U**ster and em**BE**d)
- c) Spot
- d) TrafficSpot
- e) InfoShield

'N': How to pick good names?

- a) CAE (-> Cluster And Embed)
- b) CUBE (-> **C**l**U**ster and em**BE**d)
- ~ c) Spot
- d) TrafficSpot
- e) InfoShield

English word

'N': How to pick good names?

- a) CAE (-> Cluster And Embed)
- b) CUBE (-> **C**l**U**ster and em**BE**d)
- ~ c) Spot
- d) TrafficSpot
- e) InfoShield

Carnegie Mellon

'N': How to pick good names?

- ☠ a) CAE (-> Cluster And Embed)
- ☠ b) CUBE (-> **C**l**U**ster and em**BE**d)
- ~ c) Spot
- ✓ d) TrafficSpot
- e) InfoShield

'N': How to pick good names?

- ☠ a) CAE (-> Cluster And Embed)
- ☠ b) CUBE (-> **C**l**U**ster and em**BE**d)
- ~ c) Spot
- ✓ d) TrafficSpot
- e) InfoShield

'N': How to pick good names?

- ☠ a) CAE (-> Cluster And Embed)
- ☠ b) CUBE (-> **C**l**U**ster and em**BE**d)
- ~ c) Spot
- ✓ d) TrafficSpot
- ✓ e) InfoShield

Carnegie Mellon

Recipes for ‘good names’ – part1/4

clusterEmbed TrafficLight

Carnegie Mellon

Recipes for 'good names' – part1/4

1) 'What' – not 'how'

clusterEmbed TrafficLight

Recipes for 'good names' – part1/4

- 1) 'What' – not 'how'

  clusterEmbed TrafficLight

Carnegie Mellon

Recipes for ‘good names’ – part2/4

Heraclitus

TripWire

ZX23-S

Let's take a small break

Joke

Q: What is a professor?

Let's take a small break

Joke

Q: What is a professor?

A: The person who talks while you sleep.

Let's take a small break

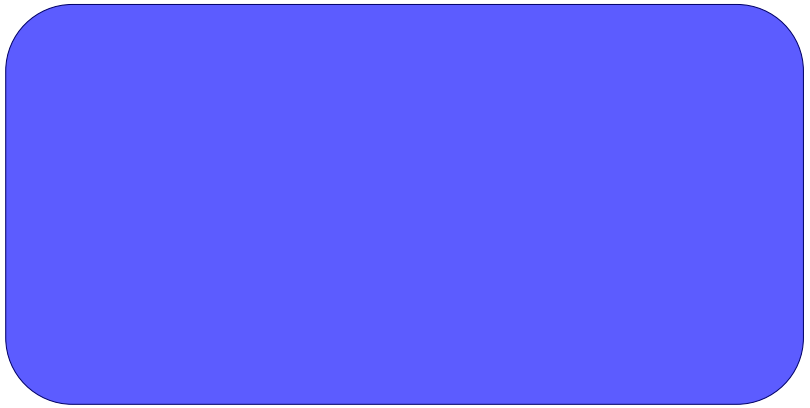
Joke

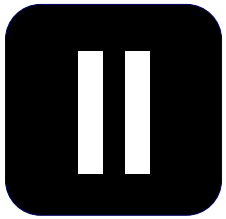
Q: What is a professor?

A: The person who talks while you sleep.

Painfully true, in my experience as a professor – sigh!

Recipes for ‘good names’ – part2/4

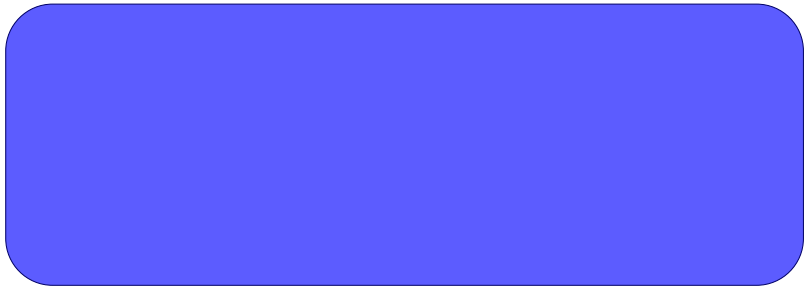


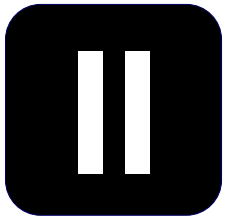


Carnegie Mellon

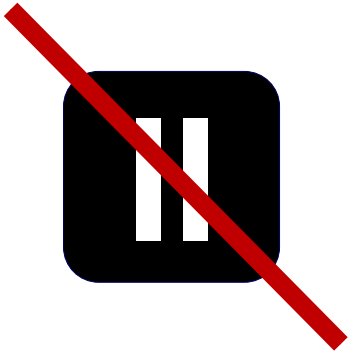
Recipes for ‘good names’ – part2/4

Heraclitus





Recipes for ‘good names’ – part2/4



Heraclitus

TripWire

Carnegie Mellon

Recipes for ‘good names’ – part2/4

Heraclitus

TripWire

ZX23-S

Recipes for 'good names' – part2/4

2) Easy to pronounce/remember

Heraclitus

TripWire

ZX23-S

Recipes for 'good names' – part2/4

2) Easy to pronounce/remember

Heraclitus

TripWire

ZX23-S

Carnegie Mellon

Recipes for ‘good names’ – part3/4

SOAR

SkySoar

Carnegie Mellon

Recipes for 'good names' – part3/4

3) But NOT an English word (google collisions..)

SOAR

SkySoar

Recipes for 'good names' – part3/4

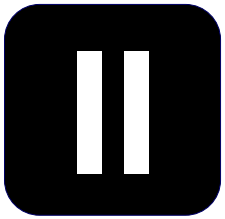
3) But NOT an English word (google collisions..)

SOAR

SkySoar

Google

Recipes for ‘good names’ – part4/4



EaglesEye

MAFIA

Poltergeist

Sherlock

Recipes for 'good names' – part4/4

4) Positive connotation

- EaglesEye
- MAFIA
- Poltergeist
- Sherlock

Recipes for ‘good names’ – part4/4

4) Positive connotation

MAFIA

Poltergeist

EaglesEye

Sherlock

Carnegie Mellon

Recipes for 'good names' – part4/4

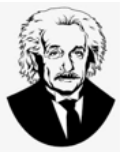
4) Positive connotation

MAFIA

Poltergeist

EaglesEye

Sherlock



Recipes for 'good names'

1) 'What' – not 'how'

clusterEmbed TrafficLight

2) Easy to pronounce/remember

Heraclitus TripWire

ZX23

3) But NOT an English word (google collisions..)

SOAR SkySoar

Google

4) Positive connotation

MAFIA EaglesEye

Poltergeist Sherlock

Carnegie Mellon

Recipes for 'good names'

1) 'What' – not 'how'


clusterEmbed TrafficLight

2) Easy to pronounce/remember

Heraclitus TripWire

ZX23

3) But NOT an English word (google collisions..)

SOAR SkySoar

Google

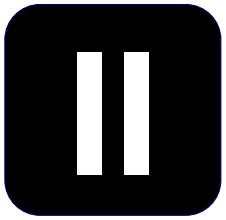
4) Positive connotation

MAFIA EaglesEye

Poltergeist Sherlock

Drill: pick a name!

Suppose that you have developed a method that detects Twitter users that exhibit similar behavior. You use random seeds, and belief propagation, to spot such groups.



- *SSTU* ('spotting similar twitter users')
- *BPSim* (Belief Propagation for similar user detection)
- *TwinSpot*

Drill: pick a name!

Suppose that you have developed a method that detects Twitter users that exhibit similar behavior. You use random seeds, and belief propagation, to spot such groups.

- ✗ *SSTU* ('spotting similar twitter users')
- ✗✓ *BPSim* (Belief Propagation for similar user detection)
- ✓ *TwinSpot*

Drill: pick a name!

✓✓• *CopyCatch* (!!) [Alex Beutel+, WWW13]

Carnegie Mellon

Drill: pick a name!

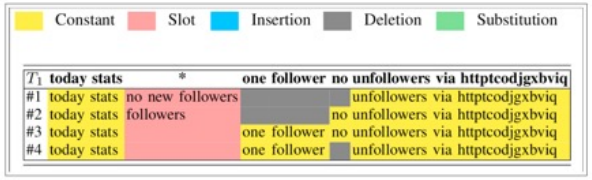
✓✓• *CopyCatch* (!!) [Alex Beutel+, WWW13]

👉 team up with a native speaker ☺

Carnegie Mellon

Summary of summary: F.A.N.

- Figure (one or more; 'what', not 'how')



- Ask a 'rhetorical' question (and answer it!)

- Name your method

CopyCatch

~~**ClusterEmbed**~~

Outline

- Top 3 lessons: 'F.A.N.'
  - Figure ('crown jewel' figure)
  - Ask (ask a rhetorical question)
  - Name (give a name to your method/system)

- A step back: 'think like a reviewer'
- More battle scars; and ~10 remedies
- Conclusions

Think like a reviewer

Reviewers:

- Pressed for time
- Un-paid / un-thanked

Q: what should authors do?

Think like a reviewer

Reviewers:

- Pressed for time
- Un-paid / un-thanked

Q: what should authors do?

- A1: **respect** & save their time (using figures, tables, **emphasis**)
- A2: attention routing

Think like a reviewer

Reviewers:

- Pressed for time
- Un-paid / un-thanked

Q: what should authors do?

- A1: **respect** & save their time (using figures, tables, **emphasis**)
- A2: attention routing

All F.A.N. recipes do that

Outline

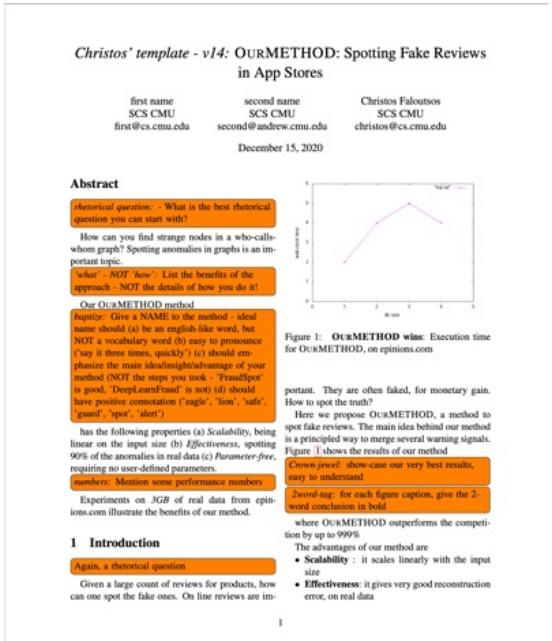
- Top 3 lessons: 'F.A.N.'
  - Figure ('crown jewel' figure)
  - Ask (ask a rhetorical question)
  - Name (give a name to your method/system)

- A step back: 'think like a reviewer'
- More battle scars; and ~10 remedies
- Conclusions

Carnegie Mellon

More recipes

- Check template tar-file with ‘orange suggestions’: https://www.cs.cmu.edu/~christos/MetaPaper
- `tar xfv; make`



Christos' template - v14: OURMETHOD: Spotting Fake Reviews in App Stores

first name, SCS CMU, first@cs.cmu.edu
second name, SCS CMU, second@andrew.cmu.edu
Christos Faloutsos, SCS CMU, christos@cs.cmu.edu

December 15, 2020

Abstract

rhetorical question: - What is the best rhetorical question you can start with?

How can you find strange nodes in a who-calls-whom graph? Spotting anomalies in graphs is an important topic.

'what' - NOT 'how': List the benefits of the approach - NOT the details of how you do it!

Our OURMETHOD method

baptize: Give a NAME to the method - ideal name should (a) be an english-like word, but NOT a vocabulary word (b) easy to pronounce ('say it three times, quickly') (c) should emphasize the main idea/insight/advantage of your method (NOT the steps you took - 'FraudSpot' is good, 'DeepLearnFraud' is not) (d) should have positive connotation ('eagle', 'lion', 'safe', 'guard', 'spot', 'alert')

has the following properties (a) *Scalability*, being linear on the input size (b) *Effectiveness*, spotting 90% of the anomalies in real data (c) *Parameter-free*, requiring no user-defined parameters.

numbers: Mention some performance numbers

Experiments on *3GB* of real data from epinions.com illustrate the benefits of our method.

1 Introduction

Again, a rhetorical question

Given a large count of reviews for products, how can one spot the fake ones. On line reviews are important. They are often faked, for monetary gain. How to spot the truth?

Here we propose OURMETHOD, a method to spot fake reviews. The main idea behind our method is a principled way to merge several warning signals. Figure 1 shows the results of our method

Figure 1: **OURMETHOD wins**: Execution time for OURMETHOD, on epinions.com

Crown-jewel: show-case our very best results, easy to understand

2word-tag: for each figure caption, give the 2-word conclusion in bold

where OURMETHOD outperforms the competition by up to 999%

The advantages of our method are

- **Scalability** : it scales linearly with the input size
- **Effectiveness**: it gives very good reconstruction error, on real data

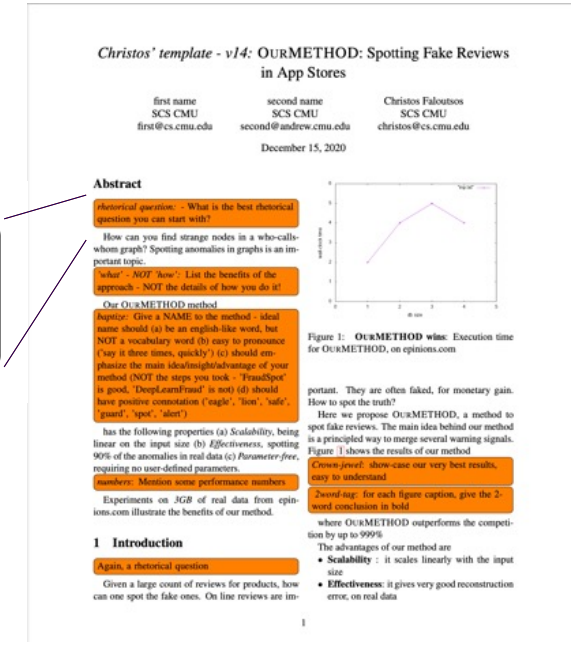
1

Carnegie Mellon

More recipes

- Check template tar-file with 'orange suggestions': https://www.cs.cmu.edu/~christos/MetaPaper
- `tar xfv; make`

rhetorical question: - What is the best rhetorical question you can start with?



Christos' template - v14: OURMETHOD: Spotting Fake Reviews in App Stores

first name, SCS CMU, first@cs.cmu.edu
second name, SCS CMU, second@andrew.cmu.edu
Christos Faloutsos, SCS CMU, christos@cs.cmu.edu

December 15, 2020

Abstract

rhetorical question: - What is the best rhetorical question you can start with?

How can you find strange nodes in a who-calls-whom graph? Spotting anomalies in graphs is an important topic.

'what' - NOT 'how': List the benefits of the approach - NOT the details of how you do it!

Our OURMETHOD method

baptize: Give a NAME to the method - ideal name should (a) be an english-like word, but NOT a vocabulary word (b) easy to pronounce ('say it three times, quickly') (c) should emphasize the main idea/insight/advantage of your method (NOT the steps you took - 'FraudSpot' is good, 'DeepLearnFraud' is not) (d) should have positive connotation ('eagle', 'lion', 'safe', 'guard', 'spot', 'alert')

has the following properties (a) *Scalability*, being linear on the input size (b) *Effectiveness*, spotting 90% of the anomalies in real data (c) *Parameter-free*, requiring no user-defined parameters.

numbers: Mention some performance numbers

Experiments on *3GB* of real data from epinions.com illustrate the benefits of our method.

1 Introduction

Again, a rhetorical question

Given a large count of reviews for products, how can one spot the fake ones. On line reviews are important. They are often faked, for monetary gain. How to spot the truth?

Here we propose OURMETHOD, a method to spot fake reviews. The main idea behind our method is a principled way to merge several warning signals. Figure 1 shows the results of our method

Crown-jewel: show-case our very best results, easy to understand

2word-tag: for each figure caption, give the 2-word conclusion in bold

Figure 1: **OURMETHOD wins**: Execution time for OURMETHOD, on epinions.com

where OURMETHOD outperforms the competition by up to 999%

The advantages of our method are

- **Scalability** : it scales linearly with the input size
- **Effectiveness**: it gives very good reconstruction error, on real data

1

Carnegie Mellon

More recipes – part 1 of 4

1. Title:
   - ✓ a. Name your method (‘what’, not ‘how’)
2. Abstract
   - ✓ a. Ask a rhetorical question
   - b. Give performance numbers

Carnegie Mellon

More recipes - part 2 of 4

3. Intro:
   - ✓ a. Again, rhetorical question
   - ✓ b. ‘crown jewel’ figure
   - c. List (bullets) 2-4 contributions
   - d. (Informal) problem definition
   - e. Give two-word summary for each contribution
4. Literature survey
   - a. ‘salesman matrix’: rows are features; columns are baselines

Carnegie Mellon

More recipes - part 3 of 4

5. Proposed method
   a. **No citations** from now on
   b. Clear problem definition
   c. Add theorems/lemmas and proofs ('QED')
6. Experiments
   a. Each sub-section should confirm each of the contributions

Carnegie Mellon

More recipes - part 3 of 4

7. Conclusions
   a. Repeat the contributions from the intro
8. Globally:
   a. ‘two-word tag’, for every figure/table caption

Carnegie Mellon

F.A.Q.

- Are these guidelines mandatory/necessary?
- Are they enough/sufficient?

F.A.Q.

- Are these guidelines mandatory/necessary?
  - **NO** (but help: + epsilon)
- Are they enough/sufficient?
  - **NO** (more, every year)

Carnegie Mellon

Resources

1. Human-trafficking detection paper:

   http://catvajiac.me/files/infoshield.pdf

   INFOSHIELD: Generalizable Information-Theoretic Human-Trafficking Detection, Meng-Chieh Lee, Catalina Vajiac, et al, ICDE 2021, Chania, Greece

2. Check 'orange suggestions' at:

   https://www.cs.cmu.edu/~christos/MetaPaper/

Conclusions – high level

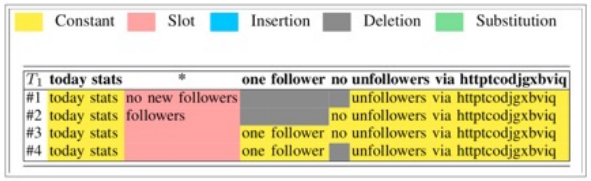
1) **Respect** reviewers' time

2) 'what', not 'how'

Carnegie Mellon

Conclusions - detailed

- F.A.N. (Figure; Ask question; Name your method)

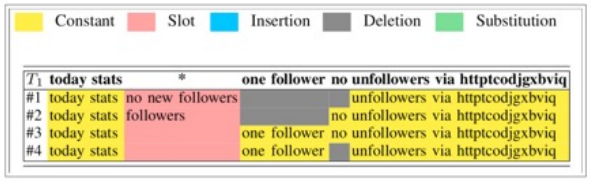


- (check orange suggestions)

Carnegie Mellon

Conclusions - detailed

- F.A.N. (Figure; Ask question; Name your method)



'ACK: … we used the MetaPaper …'

christos@cs.cmu.edu
https://www.cs.cmu.edu/~christos/MetaPaper/

Carnegie Mellon

Conclusions - detailed

- F.A.N. (Figure; Ask question; Name your method)