CHRIS JEPEWAY

MIDS 231, FALL 2016

ON THE TRAIL OF THE NETFLIX PRIZE KILLERS

A WHODUNNIT IN 7 PARTS

1. The Victim
2. The Crime
3. The Criminals
4. The Weapon
5. Crime Scene Reconstruction
6. The Case
7. The Evidence Locker

THE VICTIM

NETFLIX PRIZE

- AKA, NFP
- Home Address: http://www.netflixprize.com
- Rules: Beat Netflix's in-house recommender to win $1M
- Projected Lifespan: 2 Oct 2006 - 2 Oct 2011

THE CRIME

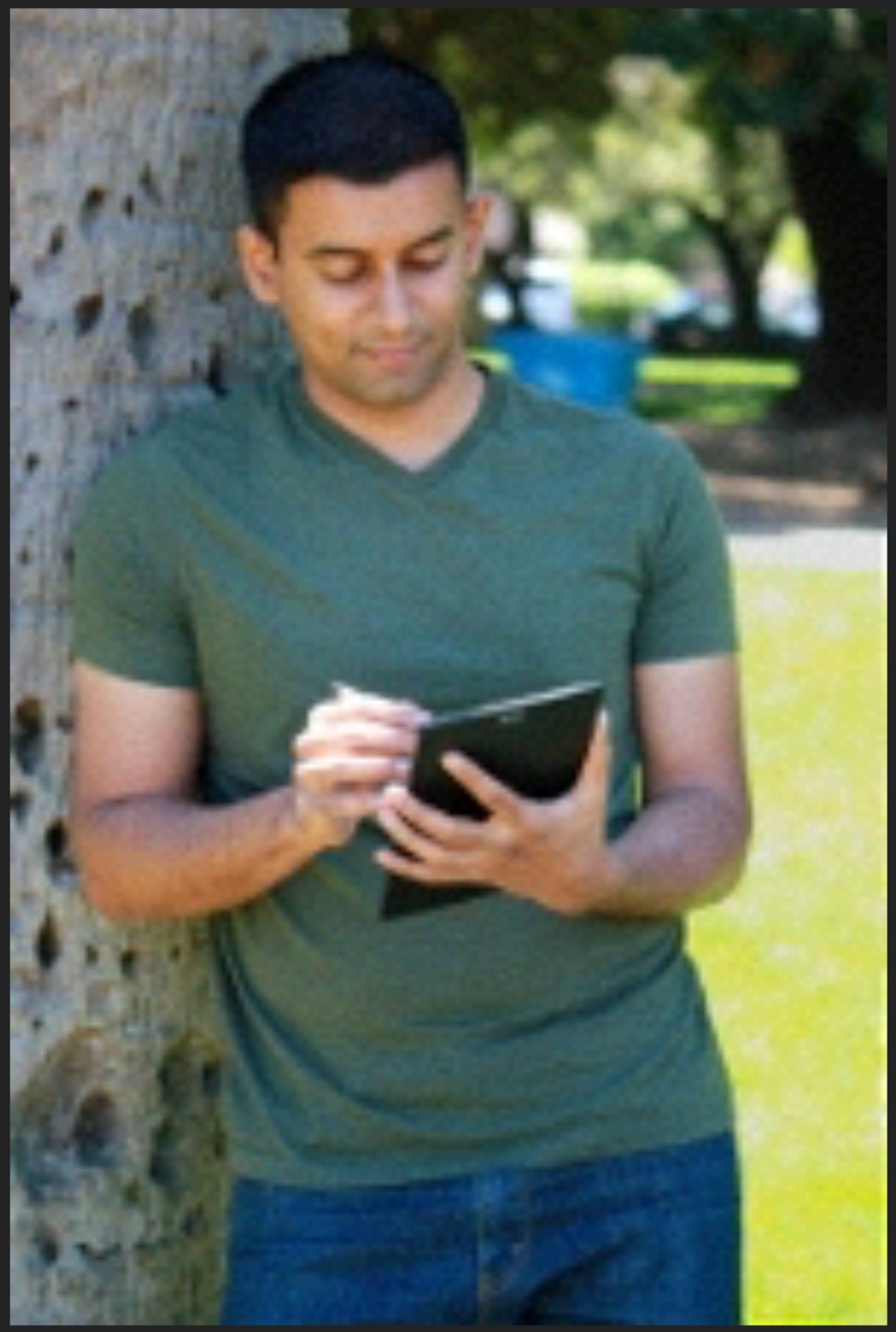
MURDER: ENDING COLLABORATIVE FILTERING…FOREVER

- NFP Fatally wounded 5 Feb 2008
- Death from complications on 21 Sep 2009
  - $1M awarded to BelKor's Pragmatic Chaos
- For collaborative filtering, nothing was ever the same

THE CRIMINALS

APB: PERP #1 – ARVIND NARAYANAN

- Known Hangouts
  - https://33bits.org
  - Freedom to Tinker
  - Princeton CS Department



APB: PERP #2 – VITALY SHMATIKOV

- Known Hangouts
  - Cornell CS Dept

THE WEAPON

ROBUST DE-ANONYMIZATION OF LARGE DATASETS (HOW TO BREAK ANONYMITY OF THE NETFLIX PRIZE DATASET)

- 2008
- Results include
  - 2 de-anonymization algorithms
  - Showing “curse of dimensionality” yields ready de-anonymization
  - Measure of how much add’l info needed when de-anonymization fails

CRIME SCENE RECONSTRUCTION

CRIMINAL RE-ENACTORS

- Two toy data sets
  - One that's been anonymized - **db**
    - **Contains sensitive data**
  - One that's loosely related to the first - **aux**
    - What the perp knows
- A handful of records
- Few fields

RE-ENACTORS: DB – ANONYMIZED CANDY DEMOGRAPHICS

Age	Home Town	Favorite Candy
12	Ghost Planet	
12	Ghost Planet	
198	Darmstadt	

RE-ENACTORS: AUX - VOTER REGISTRY

	Name	DOB	Home Town
	Brak	2/29/2004	Ghost Planet
	Sisto	2/29/2004	Ghost Planet
	Frank	1/1/1918	Darmstadt

RECREATING THE WEAPON

- mapping of comparable columns
  - Home town <=> Home town
  - Age <=> DOB
- **sim()** - compares attributes
  - 1 if hometown same, 0 if not
  - 1 if age consistent with DOB, 0 if not

RECREATING THE WEAPON: GENERAL ALGORITHM

- de_anon()

  1. Compute $S = \{ Score(aux, r') \text{ for each } r' \text{ in } DB \}$
  2. Apply match criteria over S
     1. matching set empty => output *{}*
  3. Otherwise
     1. Need a best guess? Output r' with highest score
     2. Need distribution? Output S with p.d.f

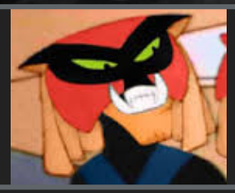
RECREATING THE WEAPON: ALGORITHM 1A

- *Score(aux, r')*
  - The min **sim()** across comparable columns
  - So, the least similar attribute counts
- Matching criteria
  - Any score in $S > \alpha$?
- Output
  - All r' with scores $> \alpha$
  - `p.d.f = U()`

RECREATING THE WEAPON: ALGORITHM 1B

- *Score(aux, r')*
  - Weighted sum across non-null aux columns of **sim()**
  - Weights are inverse of log of count of non-null column entries in *DB*
- Matching criteria
  - Are top 2 scores in *S* too close? Then, no match
  - Too close: $\Delta/\sigma_S < \varphi$
- Output
  - Top scoring record
  - $\texttt{p.d.f} = c \cdot e^{Score(aux,\ r')/\sigma_S}$

RE-ENACTING THE CRIME: ALGORITHM 1A

	Candy	Score	Prob
	?		
		1	0.5
		1	0.5
		1	0.5
		1	0.5

RE-ENACTING THE CRIME: FRANKIE'S RELAXED. HOW DID WE MISS HIM?

- The **sim()** function is too precise
  - He was born in early 1818
  - So, he's made 199 trips around the sun
  - But it's not yet 2017 (= 1818 + 199), so we say he's 198 yo
- Remember: Algorithm 1A scores on worst match

RE-ENACTING THE CRIME: BRAK & SISTO?

-  likes 

-  is all about the 

- They're fraternal twins and live together, so min over the comparable **sim()**'s is 1 for both => each is equally likely

RE-ENACTING THE CRIME: ALGORITHM 1B

	Favorite Candy	Score
		0.91
	?	
	?	

RE-ENACTING THE CRIME: AGAIN WITH BRAK & SISTO?

-  & 
- Have 2 matching attributes each
- Their top 2 are equal, so Δ is $0 < \varphi$
- No match

RE-ENACTING THE CRIME: WHAT'S α AND φ ?

- For Algorithm 1A, choose $\alpha = 1-\varepsilon$ so that de-anonymization is likely, within a given tolerance ε
- For Algorithm 1B, choose φ to reject false positives within a multiple of σ_s

CASE IS STILL OPEN

REMAINING WORK

- Complete measure of bits needed for de-identification
- Use a real data set
  - Anonymize it
  - Extract & synthesize aux
- Use two real, related data sets
- Use some of Khaled's techniques
- Contact ~~authors~~ criminals re: methods, motives
- Wrap my head around proofs

EVIDENCE LOCKER

REFERENCES

- Netflix Prize @ http://www.netflixprize.com
- Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets" @ http://ieeexplore.ieee.org/document/4531148/
- Arvind Narayanan and Vitaly Shmatikov. "De-anonymizing Social Networks" @ https://www.cs.cornell.edu/~shmat/shmat_oak09.pdf
- Khaled El Emam and Luk Arbuckle. Anonymizing Health Data - Case Studies and Methods to Get You Started @ http://shop.oreilly.com/product/0636920029229.do

ANY TIPS?