LINKSOCIAL: Linking User Profiles Across Multiple Social Media Platforms

Vishal Sharma and Curtis Dyreson
Department of Computer Science
Utah State University

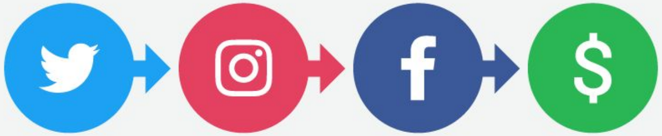
UtahStateUniversity

Outline

- Introduction of User Profile Linkage (UPL)
- Applications and challenges of UPL
- Related Work
- **LinkSocial** Framework
- Data Collection and Feature Engineering
- Predictive Model
- Computation Cost Reduction
- Results and Conclusion

User Profile Linkage (UPL)

- More than 42% of adults use more than two social media platforms
- A social media platform exposes an aspect of a user
    - Personal
    - Professional
    - Ideological
- UPL is process of linking user profiles across social media platforms
- There are several *applications* and *challenges* of UPL



Applications

- Security
- User behaviour across social media platforms
- Information verification
- Recommendation

Challenges

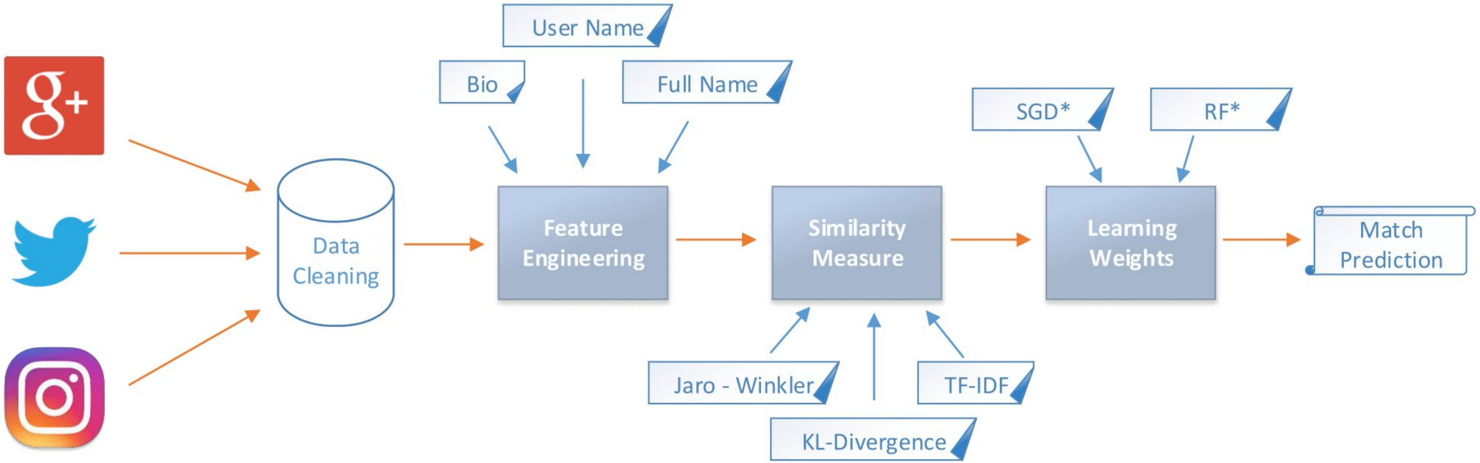
- Data collection
- Incomplete information
- False information
- Missing information across platform
- Limited access

Previous Work

COMPARISON WITH PREVIOUS RESEARCH

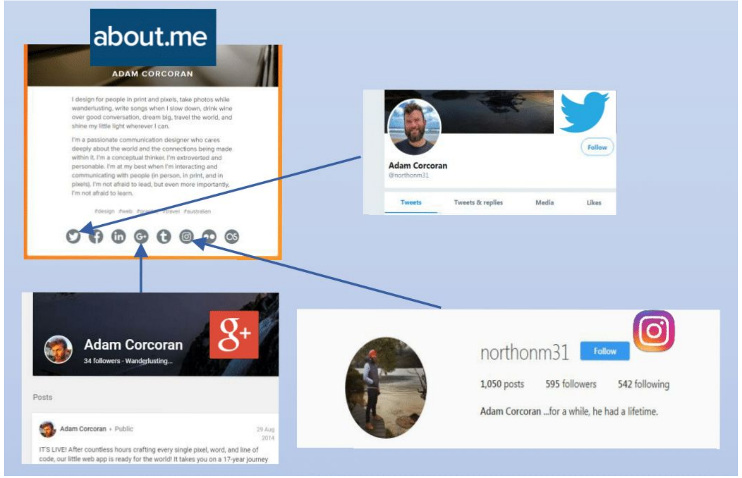
Linkage	Authors et. al	Features	Dataset Public	Scalable	Reduce Cost	Scalable Across Platform
Pair-Wise	P. Jain[28]	Private	✗	✗	✗	✗
	A. Mal[8]	Public	✗	✓	✓	✗
	R. Zafa[22]	Public	✗	✗	✗	✗
	Y. Li[31]	Private	✗	✗	✗	✗
	LINKSOCIAL	Public	✓	✓	✓	✓
Across	X. Mu[32]	Private	✗	✗	✗	✓
	S. Liu[6]	Private	✗	✗	✗	✓
	LINKSOCIAL	Public	✓	✓	✓	✓

LinkSocial Framework



Data Collection

- About.me
  - B. Lim et. al. organized 15,298 usernames
  - G+, Insta, Tumblr, Twitter, Youtube, Flickr
- We selected
  - G+, Twitter, Insta
- Public profile
  - *username*, *name*, *bio* and *profile image*
- Data collection using web crawlers



Dataset Analysis

- Avg. *username* length 11-13
- Avg. *bios* on G+ 164 characters
- Avg. *bios* on Twitter, Insta 96 and 70
- 28%, 13%, 21% atleast one missing attribute on G+, Twitter and Insta
- Lot of deactivated profiles
- 2% of Insta profile all attributes missing

Social Media	*Profile Count*
Instagram - Google+	614
Twitter - Instagram	2451
Google+ - Twitter	2974
Google+ - Twitter - Instagram	7729
	13768

Feature Engineering

- Bi-gram: Captures a range of different ways to create a *username.*
- Character distribution: bi-gram cannot capture all scenarios (e.g., john_snow, nhoj_wons)

	username	**name**	**merge**	**Similarity measure**
example	john_snow	John snow	john_snow John snow	
Bi-gram	[jo,oh,hn,n_,_s...]	[jo,oh,hn,n , s...]	[jo,oh,.. Jo,hn, ...]	Jaccard Similarity
Character dist	{j:1, o:2,h:1 ...}	{j:1, o:2, ...}	{j:2, o:4, s:2, ...]	KL Divergence

- Profile picture similarity
  - Openface crops image to extract face and represents it in 128D vector
  - Use *euclidean distance* to calculate distance between vectors

Matching

- Generate scores from engineered features
- Use Stochastic Gradient Descent (SGD) to learn feature weights
    - Mean Squared Error (MSE) as cost function
    - Learning rate of 0.001 with 1,000 iterations
- We use computed weights on test data to predict a match
    - Highest score producing profile is considered a match
- *Accuracy*: Number of correctly predicted match / Number of correct match

Computation Cost

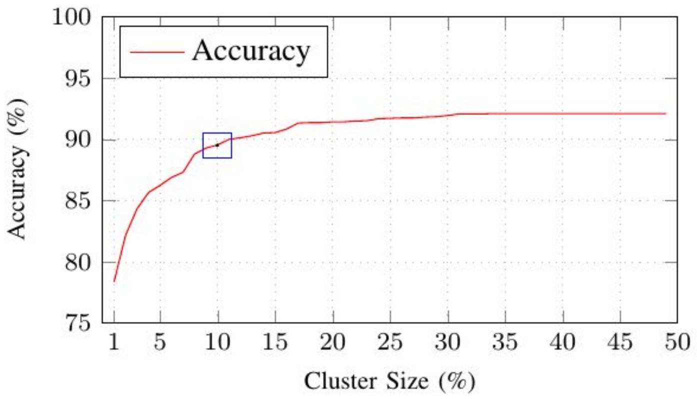
- UPL is computationally expensive
- Consider 7,729 ***pair*** user profiles (G+, Twitter)
- Number of comparisons : 7,729 * 7,729 = 59,729,712
- Assuming we perform 1,000 comparison/sec (high ballpark)
- Total time UPL ~17 hours

Imagine UPL on millions of user profiles

Computation Cost Reduction

- We introduce *candidate profile clustering*
- We choose bi-gram of *username* and *name* as profile clustering features
- We **rank** using *Jaccard Coefficient* as similarity measure

- *Cluster Size*: Percentage of top score profile
- We choose top 10% of profiles.



Experimental Setup

- Baseline
    - Jaro-Winkler: *username* and *name* similarity
    - TF-IDF and Cosine: *bios* similarity
    - Match would be highest score
    - Each feature equal weights
- Calculating Weights
    - Generate all features correct and (equally) incorrect matches
    - Using Random Forest (RF) and Stochastic Gradient Descent (SGD) for weights calculation
- Dataset randomly sampled 60-40 for Training and Testing

Results

LINKSOCIAL PERFORMANCE ON PAIR-WISE UPL

	Social Media Pairs (Accuracy %)		
Experiments	G+≡I	T≡I	G+≡T
baseline	55.36%	77.86%	56.86%
Prediction without engineered features and clustering.			
with RF	77.53%	82.08%	77.14%
with SGD	76.61%	82.21%	66.24%
Prediction with clustering, no engineered features.			
with CP & RF	82.62%	83.33%	81.40%
with CP & SGD	82.47%	83.32%	81.19%
Prediction with engineered features, no clustering.			
with RF	86.54%	91.17%	84.56%
with SGD	**86.63%**	**91.68%**	**84.58%**
Prediction with engineered features and clustering.			
with CP & RF	84.85%	87.92%	83.20%
with CP & SGD	**84.91%**	**88.29%**	**83.23%**

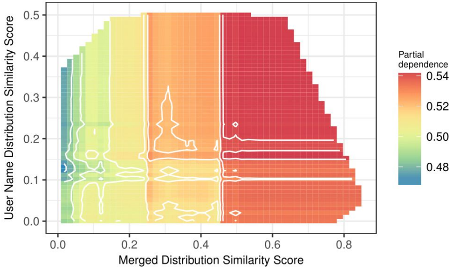
LINKSOCIAL PERFORMANCE ON MULTI-PLATFORM UPL

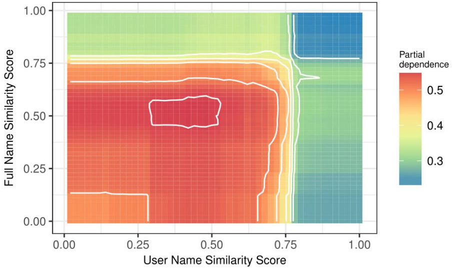
	Cross Platform		
Experiments	T→(G+, I)	G+→(T, I)	I→(G+,T)
CP & RF	71.56%	72.50%	73.70%
CP & SGD	**72.95%**	**72.86%**	**74.18%**

*RF–Random Forest, SGD–Stochastic Gradient Descent, CP–Candidate Profiles using Clustering, T–Twitter, G+–Google+, I–Instagram

Model Interpretation

- Partial Dependence plots
- *Username* dist. and *Merged* dist. highly correlated to model
- *username* and *name* similarity score until 0.75 are highly correlated
- Instances when *username* and *name* similarity scores are very high (close to 1), selected profiles do not belong to the same individual
- <u>Conclusion</u> *username* and *name* are unreliable features for linking profiles





Conclusion

- We investigate problem of UPL (pair and across platform)
- We proposed a solution
    - Linksocial: A Large scale, Efficient and Scalable
- We perform data collection, feature engineering and train model
- We perform extensive experiments and evaluation on LinkSocial
- We achieve
    - **91.68**% in *pair-wise* linkage
    - **74.18**% in *multi platform* linkage

Future Work

- Analyse user behaviour across platform
- Adding more features (text, videos, images)
- Enhance Scalability and Efficiency of LinkSocial
- Evaluate LinkSocial on more social media platforms

Vishal Sharma

Utah State University
vishalusu@gmail.com

Contribution

- *Data Collection*: Across social media platform user
- Feature Engineering
- Computation cost reduction
- Predictive Model
- Model Analysis