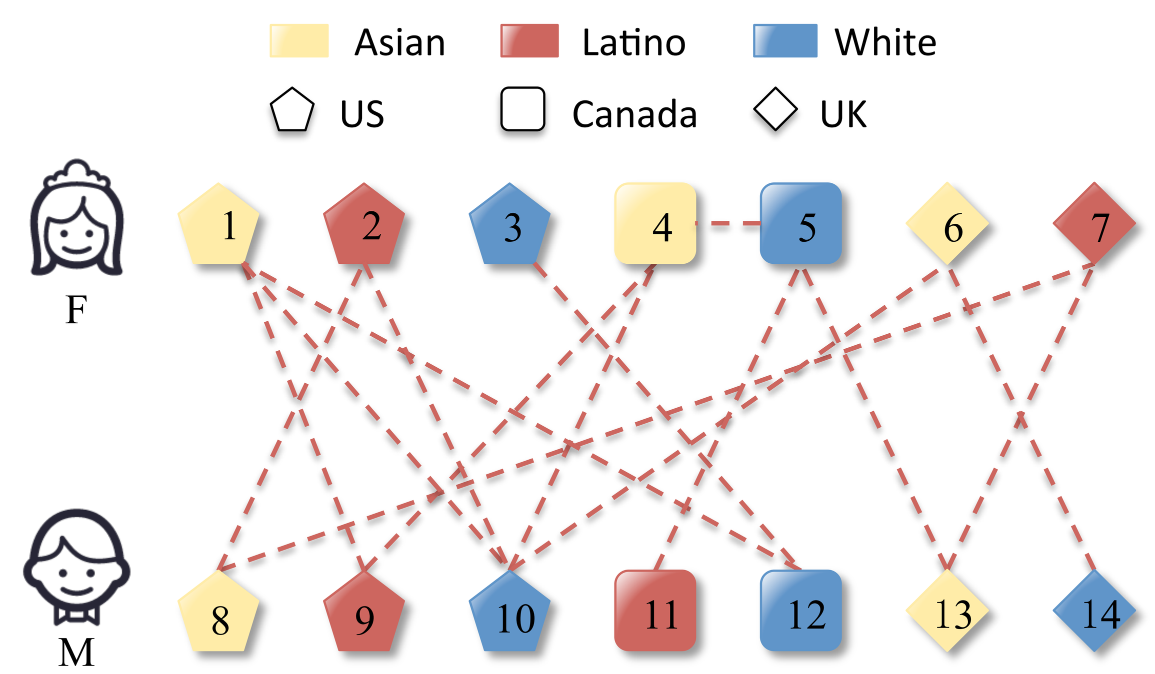
Mining Social Ties Beyond Homophily

Hongwei Liang, Ke Wang
Simon Fraser University

Feida Zhu
Singapore Management University

Introduction & Motivation



A Toy Dating Network with Node Attributes Info (Sex, Race, Location)

Social Ties (Group Relationships) $l \xrightarrow{w} r$

Leverage both graph topology and attributes information

R_1: (Sex: M) $\xrightarrow{dating}$ (Sex: F, Race: Asian)
conf = 7/14; supp = 7/15

R_2: (Sex: M, Race: Asian) $\xrightarrow{dating}$ (Sex: F, Race: Asian)
conf = 0; supp = 0

"All men except Asians preferred Asian women"

Homophily In Social Ties

- Homophily principle: love of the same
  - ✓ Contacts between similar people occur at higher rate
  - ✓ Homophily is attribute specific: e.g. Race : non-homophilic; Location: homophilic
- Homophily effect is **well-known** and often "**dominant**"

R_3: (Sex: M, Location: US) $\xrightarrow{dating}$ (Sex: F, Location: US)
conf = 4/6; supp = 4/15

Beyond Homophily

R_4: (Sex: M, Location: US) $\xrightarrow{dating}$ (Sex: F, Location: Canada)

standard confidence? $conf = 2/6$, not interesting	VS	new metric that remove homophily? $nhp = 2/(6-4) = 100\%$, **interesting !**

support of the homophily effect (Sex: F, Location: US) $\xrightarrow{dating}$ (Sex: M, Location: US) is 4/15

Reads as: if a female from US does NOT want her partner to be from US, there is a high chance that she prefers a partner from Canada.

- New Interestingness Metric
  - ✓ Non-homophily preference (nhp): a conditional probability that EXCLUDE "homophily"

$$nhp\left(l \xrightarrow{w} r\right) = \frac{supp(l \xrightarrow{w} r)}{supp(l \wedge w) - supp(homophily\ effect)}$$

Example: (Sex: F, Location: US) ⟶ (Sex: M, Location: Canada)
(Sex: F, Location: US) ⟶ (Sex: M, Location: US)

- ✓ Capture "secondary bonds" beyond "primary bonds"
- ✓ *nhp* does not have the regular anti-monotonicity

Problem Definition

Mining Top-k GRs

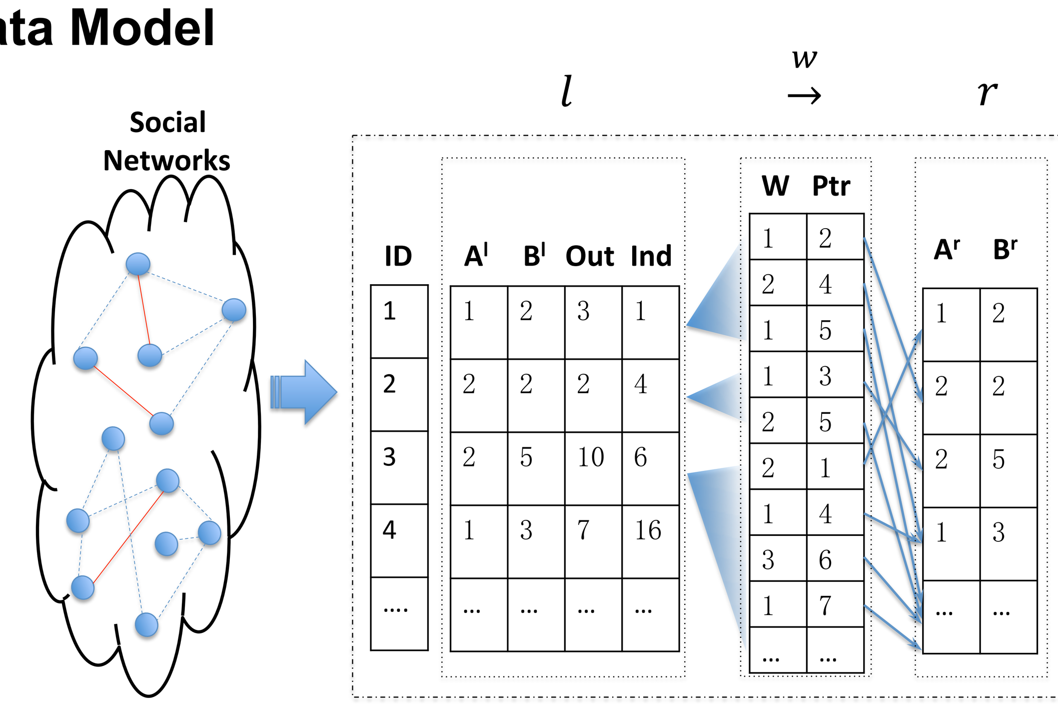
- ✓ Given: an information network, the setting of homophily for attributes, a *supp* threshold, a *nhp* threshold and an integer *k*
- ✓ Goal: discover the top-*k* interesting GRs, ranked by *nph* followed by *supp*, and each of them satisfies the *supp* and *nhp* thresholds

Solutions

Challenges

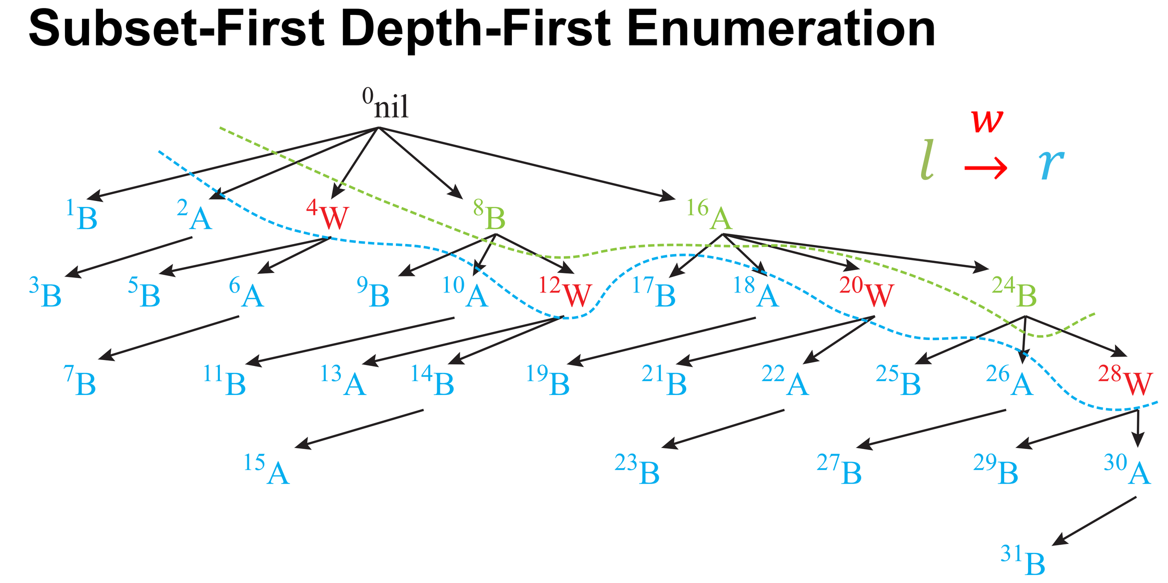
- Storage
  - ✓ Space = $|E| \times (2 \times \#Attr_V + \#Attr_E)$, if single table storage
- Computation
  - ✓ Exponential order of attributes value combination
  - ✓ *nhp* does not have anti-monotonicity
  - ✓ If only *supp* pruning: small threshold, and post-processing is needed
- How to deal with?
  - ✓ Storage: favourable data modeling
  - ✓ Computation: ingenious enumeration with efficient pruning strategies

Data Model



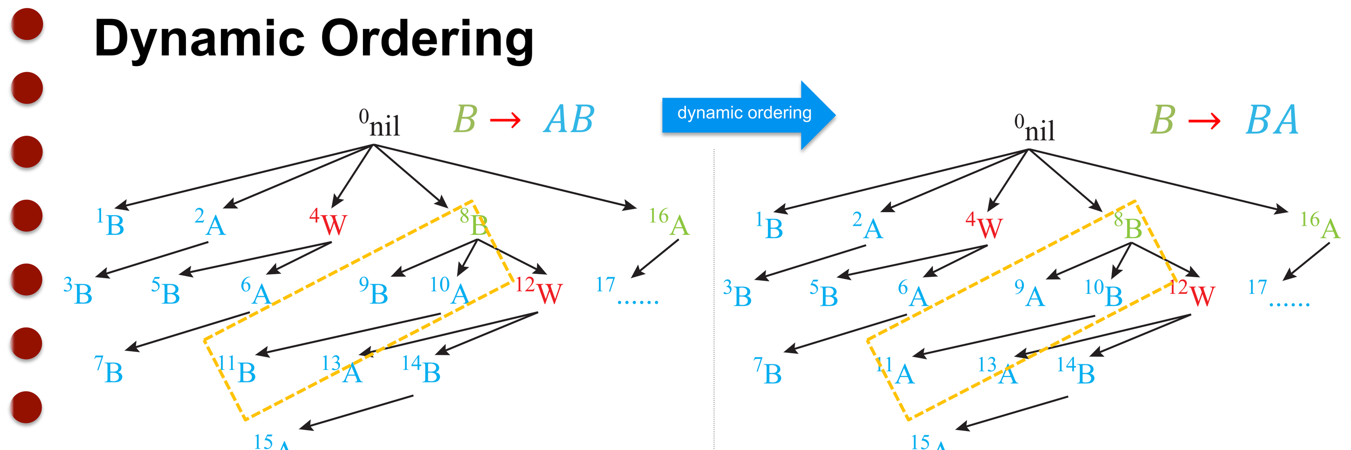
- Compact 3-table data presentation
  - ✓ Combine profile data and graph topholgy
  - ✓ No redundancy, data linked by pointers
  - ✓ Space = $|V| \times (\#Attr_V + 2) + |E| \times (\#Attr_E + 1) + |V| \times \#Attr_V$

Subset-First Depth-First Enumeration



- ✓ Subset-First: some kind of reverse order, all parts of *supp*, including that for homophily effect, are available when computing *nhp*
- ✓ Depth-First: only materialize the current branch

Dynamic Ordering



- ✓ Dynamically order the homophily attributes, on the basis of whether the same attributes were enumerated in the LHS
- ✓ $nhp\left(l \xrightarrow{w} r\right)$ for the GRs with same $l \xrightarrow{w}$ becomes anti-monotone

Multiple Pruning Strategies

- ✓ *supp* based pruning
- ✓ *nhp* based pruning
- ✓ Top-*k* pruning tights up the *nph* threshold

Experiments

Datasets

- Pokec Social Network Data
  - ✓ 1,436,515 users and 21,078,140 edges
  - ✓ 6 node attributes
- DBLP Co-authorship Data
  - ✓ 28,702 authors and 66,832 directed edges
  - ✓ 2 node attributes and 1edge attribute

Interestingness Evaluation

- Top-*k* GRs results ranked by *nhp* vs. the results ranked by standard *conf*

(a) Pokec data set

Ranked by *nhp*	Ranked by *conf*
P1: **(L:*Chat*)→(L:*Good Friend*)** $nhp = 69.5\%$; $supp = 649723$ ($conf = 30.9\%$)	(R:*27*)→(R:*27*) $conf = 72.2\%$; $supp = 250930$
P2: **(E:*Basic*)→(E:*Secondary*)** $nhp = 68.7\%$; $supp = 682715$ ($conf = 15.4\%$)	(R:*24*)→(R:*24*) $conf = 66.1\%$; $supp = 197374$
P3: **(E:*Preschool*)→(E:*Basic*)** $nhp = 66.1\%$; $supp = 54765$ ($conf = 30.4\%$)	(R:*32*)→(R:*32*) $conf = 65.1\%$; $supp = 143219$
P4: **(E:*Hardly Any*)→(E:*Basic*)** $nhp = 65\%$; $supp = 34099$ ($conf = 30.7\%$)	(R:*10*)→(R:*10*) $conf = 65\%$; $supp = 279623$
P5: **(L:*Sexual Partner*) → (G:*Female*)** $nhp = 64.7\%$; $supp = 468012$ ($conf = 64.7\%$)	(L:*Sexual Partner*) → (G:*Female*) $conf = 64.7\%$; $supp = 468012$
P207: **(G:*Male*, A:*25-34*) → (A:*18-24*)** $nhp = 50.8\%$; $supp = 593785$ ($conf = 33.9\%$)	

(b) DBLP data set

Ranked by *nhp*	Ranked by *conf*
D1: **(A:*AI*)→(P:*Poor*)** $nhp = 74.3\%$; $supp = 31330$ ($conf = 74.3\%$)	(A:*AI*)→(A:*AI*) $conf = 88.8\%$; $supp = 37458$
D2: **(A:*DB*)$\xrightarrow{often}$(A:*DM*)** $nhp = 71.5\%$; $supp = 98$ ($conf = 6.98\%$)	(A:*DB*)→(A:*DB*) $conf = 88.7\%$; $supp = 44980$
D3: **(P:*Poor*)→(P:*Poor*)** $nhp = 70.6\%$; $supp = 63174$ ($conf = 70.6\%$)	(A:*IR*)→(A:*IR*) $conf = 75.9\%$; $supp = 16020$
D4: **(P:*Excellent*)→(A:*DB*)** $nhp = 68.1\%$; $supp = 2744$ ($conf = 68.1\%$)	(A:*AI*)→(P:*Poor*) $conf = 74.3\%$; $supp = 31330$
D5: **(A:*IR*)→(P:*Poor*)** $nhp = 68.1\%$; $supp = 14368$ ($conf = 68.1\%$)	(A:*DM*)→(A:*DM*) $conf = 72.3\%$; $supp = 14232$
D16: **(A:*AI*, P:*Good*)→(A:*DM*)** $nhp = 55.2\%$; $supp = 272$ ($conf = 11.6\%$)	

- Case study
  - ✓ P5: it derives $(G: Male, L: Sexual\ Partner) \rightarrow (G: Female)$ $nhp = 68.1\%; supp = 392652$
    $(G: Female, L: Sexual\ Partner) \rightarrow (G: Male)$ $nhp = 48.8\%; supp = 71699$
    This pair suggests a big difference in the preference of opposite sex partners by males and females
  - ✓ D2: this suggests that authors in the DB area often collaborate with those in the DM area when collaborating with those not in their own area

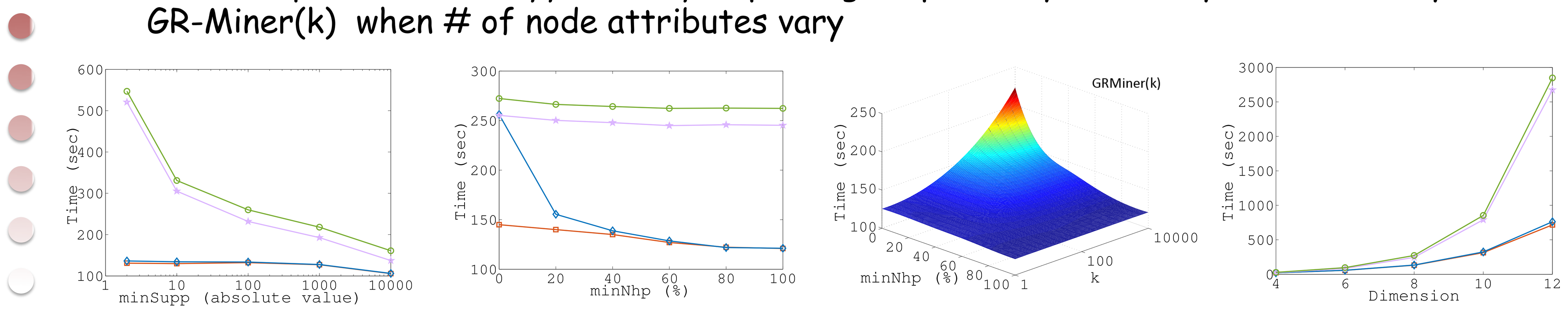
Efficiency Study (running time)

- Properties of algorithms

A: *supp* based pruning B: compact 3-table data storage C: *nhp* based pruning D: top-*k* pruning

GRMiner(k) A+B+C+D; GRMiner A+B+C; BL2 A+B; BL1 A

- Test the power of *minSupp*, *minNhp*, *k* pruning respectively and study the scalability of GR-Miner(k) when # of node attributes vary



(a) Time vs $minSupp$ (b) Time vs $minNhp$ (c) Time vs k and $minNhp$ (d) Time vs dimensionality

Conclusion & Future Work

Conclusion

Understanding how individuals form connections in a social network holds the key in many emerging applications. The literature primarily focused on the connections resulting from the homophily principle observed on social ties. In this work, we took a step in the direction that how to extract "novel" connections that are not expected from homophily by modeling the impact of homophily in the interestingness measure of connections. We formulated this problem as mining top-*k* group relationships from a social network and presented an efficient solution. This work is helpful in user behavior analysis, friend/products recommendation, missing value inference, etc.

Future work

- ✓ Alternative metrics other than *nhp*
- ✓ Deal with unstructured data
- ✓ Predictive model

ICDE 2016
May 16-20, 2016 · Helsinki, Finland