マッチングアルゴリズムを用いた匿名化手法について

宇野 毅明　情報研 & 総研大
村上 啓介　情報研

動機

・k 匿名化とは、データのレコードの各項目が、「他人にばれる属性に限定すると、k-1個の他のレコードと見分けが付かない」という状態になるよう、データを一般化する物である

・ 現実的には、k個以上のレコードがまったく同じになるように一般化することで、これを達成している　← モデルと定義が乖離している

・ 問題の定義を正確にインプリして匿名化すると、より損失の少ない匿名化ができるだろう

zip	sex	income	hobby
1??	M	1,000	tennis
?1?	F	2,000	tennis
?1?	F	1,000	TV
12?	F	1,000	tennis
1??	M	2,000	football
12?	F	2,000	TV
1??	M	1,000	football

zip
111
211
112
121
122
122
121

zip	sex	income	hobby
1?1	M	1,000	tennis
?1?	F	2,000	tennis
1?2	F	1,000	TV
??1	F	1,000	tennis
122	M	2,000	football
12?	F	2,000	TV
1??	M	1,000	football

定義:
データベースDの任意のレコード r に対し、Dを一般化したデータベース D' の k 個のレコードが、r の一般化になっているなら、D' は k 匿名化という

問題: **データベース D に対して、最小コストの匿名化 D' を求めよ**

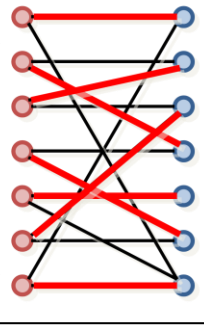
自然な定義の欠陥

・ 元データのレコードと、一般化したデータのレコードで、対応できる物間に枝を引くと2部グラフができる

・ 元データの各レコードの次数が k 以上なら、k 匿名化になっている
　しかし、、、

・ このグラフの、完全マッチングが一つのレコード間の対応になる
　➔ マッチングに必ず使われる枝は「ばれている対応」
　マッチングに使われない枝は「対応しないレコード」（実際にはない）
こういう枝は取り除いてから次数を評価したい

name	zip code	sex
Alan	111	M
Bettina	211	F
Christina	112	F
Devola	121	F
Edmond	122	M
Flora	122	F
Georgia	121	M



zip code	sex	income	hobby
1?1	M	1,000	tennis
?1?	F	2,000	tennis
1?2	F	1,000	TV
??1	F	1,000	tennis
122	M	2,000	football
12?	F	2,000	TV
1??	M	1,000	football

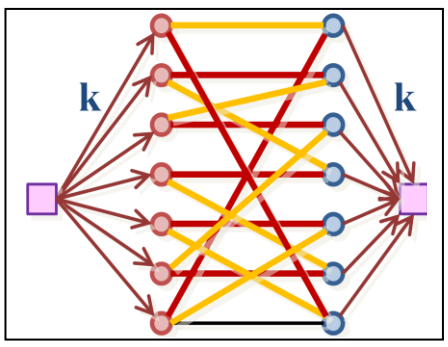
・「2部グラフが独立した完全マッチングを幾つ含むか」で、匿名度を評価する ➔ マッチング匿名性、と呼ぶ （最小次数 ≠ マッチング匿名性）

・ 残念ながら、最適化問題はNP完全

匿名度の計算

・マッチング匿名性は、与えられた一般化データベースの匿名化を判定することも自明でない

・実は、最大流問題を解くことで計算できる
右のネットワークで、流量 k を流せたら
k 匿名性がある。計算時間は $O(n^2m)$ 時間



・多項式時間でできる、とは言ってもデータが巨大なので、もっと実用的に速い方法が欲しい

最小コスト匿名化

・2マッチング匿名性なら、多項式時間で解ける
　（自明な対応＋最小重みマッチング、で最適解が得られる）
　（枝の重みは、対応関係になるために必要な一般化（？にする）属性の数

・3以上では、最小重みマッチングなどの方法ではうまくいかない
　← あるレコードを他の2つのレコードに対応させるとき、？にする属性の数は（共有されて）枝の重み和より少ないかもしれない

・実用的には、最小費用流で求めた解はいい近似解（あるいは初期解）になりそうだ（改善もしやすそう）

・初期解が有れば、近傍探索型の近似解法が動きそう
　← マッチングの付け替えは、負重みの交互サイクル（マッチングの枝と宋でない枝が交互に現れるサイクル。順番にたどりながら探索して簡単に見つけられる）を見つけて枝を入れ替えることで、少し良い解に移動できる

計算実験途中

・ まず最初のフェイズ、最小費用流が大変
　← データがものすごく大きいので、2乗かかるアルゴリズムはほぼ動かない
　➔ 最短路繰り返し法を試してみたが、パフォーマンスは今ひとつ
　（近似解ならある程度高速に得られる）

・ ここに関しては、Goldburg さんのpreflow push型のいいアルゴリズムがあるので、それに期待

・ つぎに、負重みサイクルの発見が大変。ベルマンフォード法をまともに動かすといつまでたっても終わらないので、なんとかさぼる方法を考え中

・ もっと簡単な方がいいかもなー、とか思っています

・ さらに、データの入手ができない。人工的に生成したデータを使ってますが、これって果たしていい物なのか、、、

・ でも、そもそも、基準となるランダムデータのような物も、分野として、絶対に必要なので、どのように生成するべきなのか、世の中のデータはどんな感じで、どういうところが計算上難しいのか、という知見を得ることも大事

・ まだまだ先は長いです。