Odkrivanje goljufij na osnovi analize socialnih mrež

Lovro Šubelj

Diplomsko delo

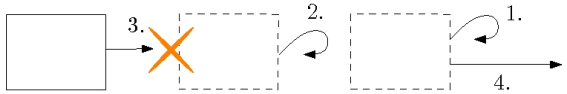
Mentor: doc. dr. Marko Bajec
Somentor: doc. dr. Matjaž Kukar

September 2008

Avtomobilske goljufije

- izsiljene ali uprizorjene prometne nesreče – lažne poškodbe

- problemi
- skupne značilnosti
- obstajajo znane sheme za uprizarjanje nesreč



- posebej zanimive so organizirane skupine posameznikov
- **Cilj:** odkrivati predvsem take skupine (ne posamezne nesreče!)

Vhodni podatki

- občutljivi podatki, navadno so neoznačeni
- v sistemu uporabimo zgolj podatke iz policijskih zapisnikov o nesrečah (pogosto v praksi):
  - entitete: nesreče, udeleženci, vozila, policisti
  - statični podatki (atributi): sumljivost nesreče, čas, gmotna škoda, poškodbe; spol, starost udeležencev
  - relacijski podatki: voznik-nesreča; sopotnik-vozilo; policist-nesreča . . .
  - ostali podatki v večini ne nosijo nobene informacije

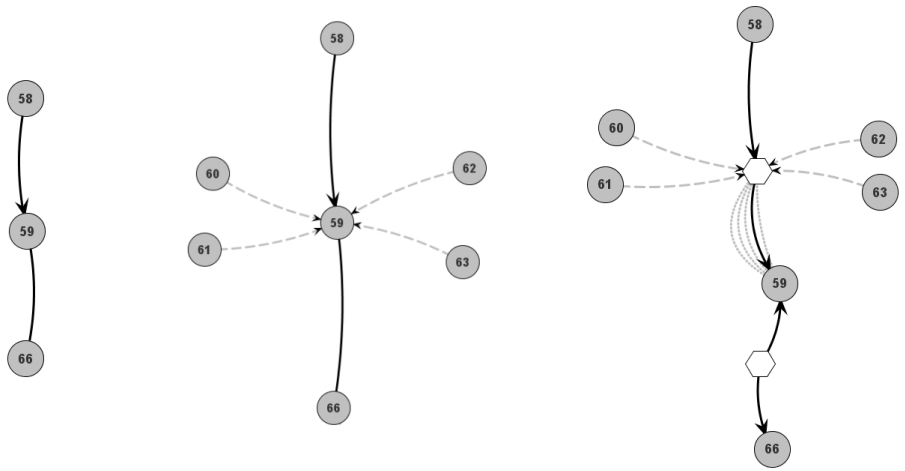
Oris sistema

Sistem razdelimo v tri dele:

1. (relacijske) podatke predstavimo s pomočjo mrež. Slednje po potrebi tudi nekoliko poenostavimo
2. identificiramo sumljive (povezane) komponente v mreži, ostale zavržemo
3. poiščemo ključne entitete v vsaki sumljivi komponenti – te predstavljajo goljufivo skupino

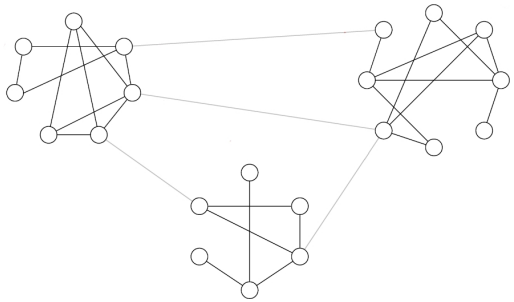
1. del: predstavitev z mrežami

- (relacijske) podatke predstavimo s pomočjo mrež: vozlišča ustrezajo različnim entitetam, povezave pa relacijam med njimi
- mreže omogočajo formulacijo kompleksnih relacij – težko dosegljivo z običajno atributno predstavitvijo podatkov
- ustvarimo različne vrste mrež: mreža voznikov, sopotnikov, nesreč



1. del: predstavitev z mrežami - nadaljevanje

- mreže po potrebi tudi nekoliko poenostavimo
- vsako povezano komponento mreže delimo glede na skupnosti, ki se pojavljajo znotraj nje (brez izgube za splošnost)

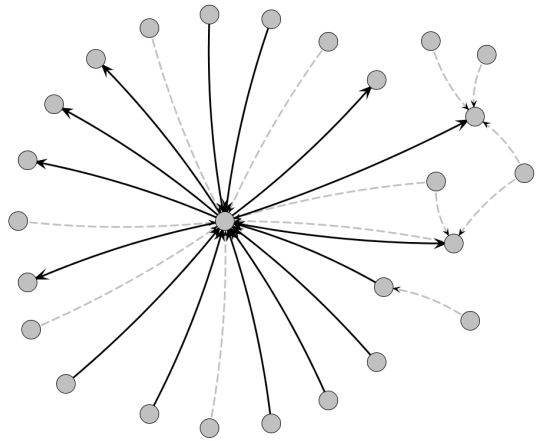


- postopek: rekurzivno odstranjujemo povezave z največjo vmesnostjo

$$B(e) = \frac{|\{(v_i, v_j)|v_i, v_j \in V \wedge i < j \wedge e \in g(v_i, v_j)\}|}{\binom{n}{2}}$$

2. del: identifikacija sumljivih komponent

- izpostavimo lastnosti komponent, ki ustrezajo goljufivim skupinam: razmerje med številom udeležencev in nesreč; skupna sumljivost nesreč; premer komponente; največja stopnja in največja vmesna centralnost vozlišča



- kot sumljive identificiramo komponente, ki po večini lastnosti izstopajo

3. del: odkrivanje ključnih entitet

- **Ideja:** vsak udeleženec je dobro opredeljen s svojimi nesrečami, vsaka nesreča je dobro opredeljena s svojimi udeleženci (ter seveda tudi s svojimi statičnimi lastnostmi)
- ideja se ujema z relacijo sosednosti v mreži nesreč
- postopek: sumljivost vozlišča je enaka uteženi linearni kombinaciji sumljivosti njegovih sosedov
- z iterativnim ocenjevanjem premagamo lokalnost
- ocenjujemo zgolj sumljivost udeležencev, ostale entitete ocenimo iz teh

3. del: odkrivanje ključnih entitet - nadaljevanje

Metoda:

1. inicializiraj sumljivost udeležencev
2. dokler $\sum_{i=1}^{s}(s^k(u_i) - s^{k-1}(u_i))^2 > \epsilon^2$, ponovi za $\forall i$:

$$s^{k+1}(a_i) = f_{ent}(a_i) \sum_{e=\{v,V_K(a_i)\}\in E(K),\ x=V_K^{-1}(v)} f_e(e,a_i)\ s^k(x)$$

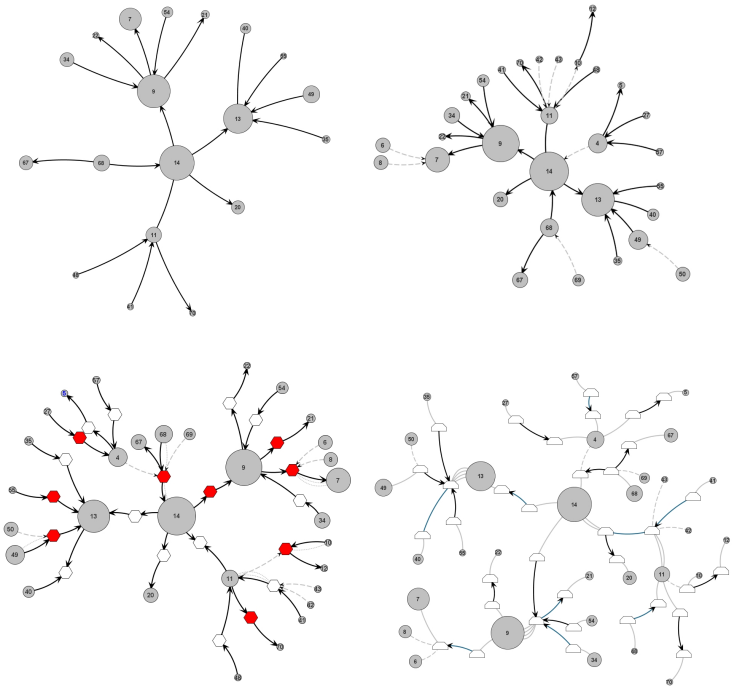
$$s^{k+1}(u_i) = \gamma s^k(u_i) + (1-\gamma) f_{ent}(u_i) \sum_{e=\{v,V_K(u_i)\}\in E(K),\ a_j=V_K^{-1}(v)} f_e(e,u_i)\ s^{k+1}(a_j)$$

$$s^{k+1}(u_i) = \frac{s^{k+1}(u_i)}{\sum_{j=1}^{s} s^{k+1}(u_j)}$$

3. normaliziraj sumljivost udeležencev glede na skupno sumljivost nesreč

Predstavitev, uporaba znanja

- popolnoma avtomatski sistem v praksi ni mogoč
- rezultate na koncu prikažemo analitiku – uporabimo mreže
- uporaba znanja pridobljenega z raziskavo



Rezultati testiranja

- testni nabor podatkov: 40 nesreč, 71 udeležencev (47 voznikov), 48 policistov, 68 vozil ...
- podatki niso izbrani naključno – kar 34% udeležencev je označenih za goljufe
- pri testiranju izpostavimo vse udeležence z nadpovprečno sumljivostjo – dosežemo $AUC = 83.87\%$ (podobno CA, priklic, specifičnost)
- pazljivost pri interpretaciji rezultatov:
  majhen nereprezentativen vzorec; uporabljamo ga pri zasnovi sistema (a ne za učenje!); vprašljivost pravilne označitve podatkov

Sklep

- z doseženim smo zadovoljni
- veliko možnih izboljšav:
  - označeni podatki: za odkrivanje sumljivih komponent uporabimo kar neko metodo strojnega učenja; podobno za faktorje $f_{ent}(\cdot)$
  - več podatkov: vključimo še druge entitete
  - uporaba hipergrafov (hipermrež)
  - drugo