Predikcija terorističkih napada u okviru socijalnih mreža

Dejan Marković

Elektrotehnički fakultet, Beograd, Srbija

Sadržaj – *U ovom radu je opisan predlog za rešavanje ozbiljnog problema u današnjem modernom svetu, vezanom za sprečavanje terorističkih napada. Generalno problem je što su teroristi postali „pametni“ pa su počeli da koriste alternativne načine komunikacije i nije moguće klasično nadgledanje sadržaja koji se razmenjuje između ljudi. Takođe dodatni problem predstavlja I činjenica da se obično koristi i nekakvo šifrovanje osetljivog sadržaja. U ovom radu je opisan jedan predlog kako bi mogao navedeni problem da se reši.*

1. UVOD

Generalno u modernom dobu pojavio se problem nekontrolisanih vidova komunikacije, kao i mnoštvo nelegalnih tipova sadržaja koji putuju takvim vidovima komunikacije. Detekcija sifrovanog sadržaja i klasično nadgledanje sa okidačima nelegalnih reči ovde nije razmatrana mada može da se koristi kao pomoćni modul za navedenu tehniku.

Za dati snimak grafa društvene mreže, postavlja se pitanje da li možemo zaključiti koje nove interakcije među potencijalnim prekršiocima će se verovatno pojaviti u bliskoj budućnosti? Problem se formalizuje kao pod problem link-prediction problema, a mi razvijamo pristupe povezivanja predviđanje na osnovu mera za analiziranje blizine čvorova u mreži . Eksperimenti na velikim radovima prethodnih autora na istu temu ukazuju na to da se informacije o budućim odnosima biti izvađene iz same topologije mreže , i da prilično suptilne mere za otkrivanje čvora blizinu može nadmašiti više direktne mjere.

Tokom nekoliko prethodnih godina, znatna količina pažnje je posvećena analizi računarskih društvenih mreža - struktura čiji čvorovi predstavljaju ljude ili druge entitete ugrađene u društveni kontekst , i čije ivice reprezentuju interakciju, saradnju, ili uticaj između entiteta. Prirodni primer društvenih mreža obuhvata skup svih korisnika u određenoj prijateljskoj grupi, sa ivicama koje reprezentuju medjusobnu komunikaciju izmedju korisnika; skup svih zaposlenih u velikoj kompaniji , sa ivicama spajaju parove koji rade na zajedničkom projektu; ili zbirka poslovnih lidera, sa ivicama spajaju parove koji su zajedno služili na korporativnom upravnom odboru . Povećana dostupnost velikih, detaljnih struktura podataka koji reprezentuju takve mreže (kao na primer grafova) je podstakla opširnu studiju njihovih osnovnih mogućnostii primena, kao i podataka odgovarajućeg formata koji se mogu upotrebiti u popunjavanje takve strukture.

Društvene mreže su veoma promenljivi objekti (menjaju se jako brzo i kroz vreme i korz prostor); rastu u formi izmena i dodavanja novih ivica, direktna posledica pojave novih interakcija i promenu postojecih leže u osnovi društvene lestvice . Identifikovanju mehanizama pomoću kojih oni evoluiraju je i dalje problem zato sto ne postoji dovoljno sličan matematički model koji bi ga simulirao, tj. ne postoji dovoljno kompleksan model koji bi detaljno prikazao sve faktore koji dovode do evolucije događaja, i ovaj problem je bio direktna ideja za rad koji je prezentovan ovde. Problem očigledno još uvek nije moguće resšiti matematički, ali uvek je moguće pribeći nekoj heuristici, i logičan nastavak rada je data mining proces.

Sada je već moguće i formulisati sam problem u skladu sa navedenim predlogom rešenja, odnosno za dati graf stanja u socijalnoj mreži potrebno je predvideti nove veze u grafu kao i promene postojećih. Ono što bi bilo idealno je i da se navedena obrada radi brzo i u diskretnim periodima, i to naravno bez ljudske interakcije. (ili sa što manje) Generalno data-mining princip se zasniva na dva perioda tokom razvoja. Prvi zahteva formiranje modela i treniranje algoritma nad realnim (ali ipak test podacima).

Ono što je bitno jeste da ti podaci budu dovoljno reprezentativni i raznoliki da bi ovo mogli da uradimo dovoljno dobro kako bi iduća faza mogla dobro da radi svoj posao. Dodatno poboljsanje ovog sistema bi bilo da se sistem samostalno modifikuje na osnovu grešaka koje bi isto sam morao da uoči. Poslednje navedeni problem predstavlja osnovu problema veštačke inteligencije i on je već dovoljno kompleksan i bez postojećih problema tako da ćemo se mi fokusirati samo na problem koji smo prvobitno opisali, i pokušavaćemo samo da za trenutno stanje odredimo tražene podatke.

2. POSTOJEĆA REŠENJA I PROBLEMI

Ovaj rad nije prvi na temu ove metodologije za rešavanje ovog problema. Većina postojećih rešenja se samo bavi pronalaženjem jedne (odgovarajuće metodologije) Takvi načini su davali solidne rezultate, ali ne dovoljno dobre da bi se podstakao dalji razvoj takvog sistema. Detaljna analiza postojećih perfomansi i predlozenog rešenja će biti diskutovani u narednim poglavljima. Globalno gledano perfomanse prethodnih sistema su bile ispod 50% sa mnogo grešaka u detekciji kritičnih korisnika. Predlog koji se iznosi u ovom radu je hibridno rešenje koje daje dosta bolje rezultate po pitanju broja pogodaka i grešaka koje pravi usput.

3. PREDLOG REŠENJA PROBLEMA

Neki od problema su već kroz tekst spomenuti, ali detaljnije će ipak biti obrađeni u ovoj sekciji. Prvi problem koji se nameće jeste pristup informacijama. Za ovako definisani problem morali bi imati potpuni pristup svim podacima iz baza/servera same socijalne mreže što može biti problem po pitanju privilegije. Takodje pitanje je i koliko bi prostorno takvi podaci zauzimali. Generalno za potrebe data mining algoritma potrebno je sortirati podatke o samim korisnicima i postojećoj komunikaciji u nekakvu bazu i normalizovati te podatke. Sama normalizacija i formiranje baze bi trajali dosta vremena, ali tek onda bi sledio pravi posao. Potrebno je korišćenje nekog softverskog alata za data-mining kao i izbor odgovarajućeg metoda za formiranje zavisnosti u okviru samog modela.

Verujemo da primarni model koji bi uticao na rešavanje ovog problema je u oblasti industrije - evolucije modela mreža. Iako je bilo širenje takvih modela u poslednjih nekoliko godina (naročito u periodu od 2000 - 2004 godine); Ti modeli evolucije mreža se zasnivaju na replikaciji odnosa i struktura u okviru samih mreža. Što se tiče odnosa sa predikcijom veza tu se stvari komplikuju jer se menja svrha iz korena. Problem predviđanja veza takođe je vezan za problem otkrivanje nedostajućih veza u skupu podataka gde iste nisu poznate: Postoje radovi koji grade mrežu interakcije na osnovu raspoloživih podataka, a zatim pokušavaju da zaključe dodatne veze, koje ne postoje u vizuelnom modelu, ali verovatno će se pojaviti u nekom momentu. Ova linija rada razlikuje od našeg problema u samoj formulaciji problema po tome što radi sa statičkom slikom mreže, a ne uzima se u obzir evolucija mrežnog modela; takođe teži da uzme u obzir specifične osobine čvorova u mreži umesto određivanja mere povezanosti čvorova, koje se rade čisto na osnovu grafovske strukture.

Socijalna mreža se predstavlja grafom G={V, E} pri čemu je V skup čvorova koji postoje u grafu, a E skup veza odnosno interakcija koje postoje između članova. Skup e={u, v} predstavlja interakciju između u i v uspostavljenu u određenom vremenu t(e). Mi imamo zapamćenu svaku interakciju između u i v u vidu paralelnih ivica, sa potencijalno različitim vremenskim oznakama kreiranja. Za dva puta t i t', neka G [t , t'] koji označavaju podgraf od G koja se sastoji od svih ivica sa vremenskom oznakom između ta dva vremena. Problem rešavamo na sledeći način: algoritmu dajemo podgraf koji sadrži trenutno stanje korisnika i njihove prethodne interakcije u odgovarajućem vremenu, kao i vremena za koja se prediktuju nove veze u sistemu.

Saradnja se definiše kao veza između dva korisnika koji su imali bilo kakvu interakciju. Na slici ispod prikazana je analiza problema, vezanih za link-prediction problem. Na slici su navedeni podaci poput broja korisnika koji su analizirani, broj postojećih veza, kao i promena koja je predvidjena sa odgovarajućim algoritmom. Akronimi su vezani za odgovarajuće oblasti iz kojih su izvađeni podaci. Astro-ph je astrofizika, cond-mat je kondenzovana materija, i sl. Takođe bitan faktor za odabir podataka je bio bar tri rada da ima svaki autor u čvoru.

	Training Period			Core		
	Authors	Articles	Collaborations[a]	Authors	E_{old}	E_{new}
astro-ph	5,343	5,816	41,852	1,561	6,178	5,751
cond-mat	5,469	6,700	19,881	1,253	1,899	1,150
gr-qc	2,122	3,287	5,724	486	519	400
hep-ph	5,414	10,254	47,806	1,790	6,654	3,294
hep-th	5,241	9,498	15,842	1,438	2,311	1,576

Slika 1. Prikaz statističkih podataka u pojedinačnim oblastima algoritama

U ovom radu se predlaže hibridno rešenje, tj kombinacija više algoritama pri čemu se koeficijenti računaju dinamički. U ovu metodu bi ukombinovali i više data-mining algoritama, poput neuralnih mreža, K-nearest neighbour, stabla odlučivanja i slično. Hibridnost obezbeđuje najbolje moguće perfomanse po pitanju pogodaka, a dinamičnost koeficijenata omogućava prilagodljivost modela različitim grupama podataka.

Deo pristupa u lin-prediction metodama koje već postoje se zasniva na ideju da su dva čvora koji imaju preklapajući skup atributa/osobina slični i time je moguća interakcija. (Možda se ona već dogodila i u prošlosti čime se samo povećala verovatnoća da se opet ostvari) Algoritmi koji će biti spomenuti u narednom izlaganju će se zasnivati na ovoj činjenici delom, a neki i u celini.

Novina koja se prezentuje je poboljšana tehnika clustering-a (grupisanja čvorova). Naime tehnika se zasniva na metodu sličnom koji je i Edsgar Dijkstra predložio za skraćivanje putanja u grafu, s tim sto bi se pronadjene putanje upisivale u graf a sve ostale veze eliminisale. Takođe bi se i jako povezane komponente grupisale u jedan čvor čime bi se kompletna procedura jako ubrzala i približila real time cilju kome težimo.

Što se tiče dodatnog ulaska u informacije na čvorovima, proveravao bi se I sadržaj rečenica kao I sličnost sa postojećom bazom podataka (što bi naravno bio dodatni data-mining projekat nezavisan od ovog, ali psotoji već dosta gotovih rešenja)

Koristili bi neke od postojećih algoritama čiji su kvaliteti prikazani u tabeli 1 za samo jednu naučnu oblast. Tako, oni se mogu posmatrati kao izračunavanje meru blizine ili sličnosti između čvorova x i y , u skladu sa topologijom mreže. Uopšteno , postupci su prilagođena iz tehnike koje se koriste u teoriji grafova i u analizi društveno - mreže; u velikom broju slučajeva , ove tehnike nisu dizajnirani za merenje međučvorovske sličnosti i stoga moraju biti modifikovan za tu svrhu. Tabela 1 ukratko sumira i poredi medjusobno algoritme na nivou jedne oblasti . Slična je situacija sto se tiče i drugih oblasti.

Algorithm	Formula
Graph distance	ABS(shortest path)
Common neighbours	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$
Katz	$\sum_{\ell=1}^{\infty} \beta^{\ell} \cdot \lvert \mathrm{paths}_{x,y}^{\langle \ell \rangle} \rvert$
SimRank	$\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \mathrm{score}(a,b)}{\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert} & \text{otherwise} \end{cases}$

Tabela 1. Pregled postojećih algoritama za link-prediction

Primetimo da je Katz algoritam posebno markiran, jer je on direktno vezan za naš problem, tj. To je algoritam koji je bio namenjen detekciji terorističkih napada, pre prezentacije ovog rada.

Naime, *Katz* je definisano meru koja direktno sumira preko kolekcije putanja, formirajući eksponencijalni afinitet ka kraćim putanjama (što je iz očiglednih razloga jako bitno za naš problem). Formula preko koje se računaju podaci je data u tabeli 1. a objašnjenje sledi. Deo pod apsolutnom vrednošću predstavlja sve putanje duzine L unutar kolekcije koja se posmatra od čvora x do čvora y a beta je parametar prediktora (zavisi od više faktora).

Moguće rešenje može da se zasniva na ideji da su dva čvora imaju veće šanse da formiraju vezu u budućnosti, ako skupovi njihovih suseda (x) i (y) imaju veliko preklapanje. To se u mnogome već nadovezuje na ideju koja je opisana u prethodnom tekstu, a i u globalu na data-mining algoritme koji se kombinuju. Prethodni opis kada se pretoči u formulu, dobijamo formulu koja je navedena u tabeli 1. Adamic i Adar (2003) su formirali sličnu formulu za određivanje vrednosti čvorova, na osnovu osobina koje se utvrde a koje su bitne na nivou dva čvora za koji se sprovodi račun. Da bi uradili ovo, oni su izračunali osobine sličnosti čvorova i definisali sličnost istih kao

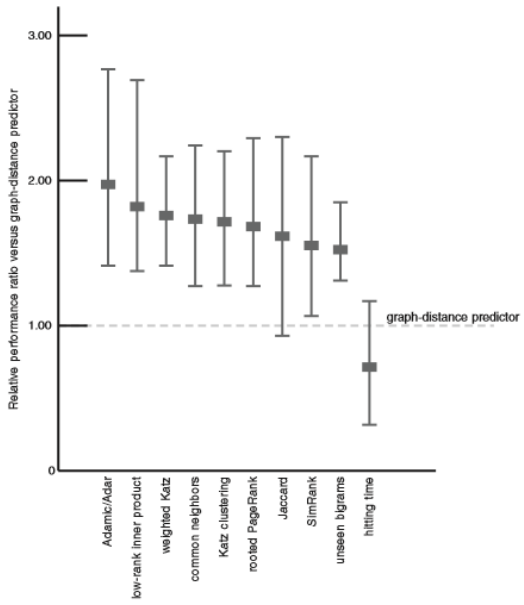
$$\sum_{z\,:\,\text{feature shared by } x,\,y} \frac{1}{\log(\mathrm{frequency}(z))}$$

Ova mera formira težinu tako što sabira zajedničke osobine dva čvora, pri čemu se ređe osobine više vrednuju.

Mogući pristup i dodatak na prethodno izlaganje jeste i sledeći princip: verovatnoća da nova veza x ima za destinacju čvor x je proporcionalna |Γ(x)|, tj. Tekućem broju suseda krajnjeg čvora. Ova metoda ne koristi činjenicu da su čvorovi moguči teroristi ali pomaze kod nekog početnog oderđivanja mogućih suseda nekom kritičnom čvoru.
U dosadašnjem izlaganju nije bilo naglašavano, ali uz početni skup podataka stanja grafa socijalne mreže, moramo imati i podatke o čvorovima koji su to kritični čvorovi po pitanju sumnje na terorizam. Takve podatke nećemo dobiti od socijalne mreže, ali takvi podaci se mogu dobiti od bezbednosnih službi zemalja. Taj skup se može dinamilki održavati sa manjim troškom od samog grafa mreže i podataka vezanih za njega, tako da će se smatrati da takav sistem postoji u okviru izlaganog rešenja.

Matrica povezanosti M može da se koristi za predstavljanje grafa saradnje G_{collab}, svi metodi koji se spominju u ovom radu polaze od matrice sličnosti M i to će takođe biti podrazumevano za ubuduće. U nekim slučajevima , ova metodologija je morala biti naglašena da je korišćena kao početni skup podatkaa (npr , u slučaju Katz tehnike sličnosti rezultata) , ali u mnogim drugim slučajevima formulacija početnog skupa podataka se nametala kao prirodna stvar. Najprostiji primer koji je prethodno naveden a koji demonstrira lakoću korišćenja matrice kao početnog skupa podataka je metod najbližih suseda: tj. Prostim prolazom kroz red odnosno kroz kolonu i analizom podataka (sumiranjem/usrednjavanjem) dolazimo jako brzo do željenih podataka. Zajednički postupak za obradu ove velike matrice M kao strukture podataka je da izabere relativno mali broj k. Zatim se izračuna rang - k matrice M. Faktor k računamo iz razloga jer najbolje predstavlja samu matricu M od svih ostalih matričnih normi. Ovaj faktor se može efikasno izračunati preko dekompozicije jednostruke vrednosti i predstavlja jezgro metoda poznatog kao latentna semantička analiza (Autori koji su pomenuli ovakav način analize su Deervester, Dumejz, Furnas, Landauer, i Harshman, 1990). Intuitivno, rad sa M_k, se može smatrati tehnikom redukcije šuma u podacima, koji je neizbežna stvar kada radimo data-mining tehnikom koja generiše većinu strukture u matrici , a takođe i uprošćava kpodatke koje treba da obradimo.

Većina navedenih algoritama funkcioniše po principu računanja nekakve težine putanje i na osnovu toga se dalje prediktuje mogućnost veze sa kritičnim čvorovima. Razmotrićemo još par algoritama radi demonstracije. SimRank je fiksna tačka sledeće rekurzivan definicija : Dva čvora su slična ako se ustanovi da su prijatelji sa sličnim susedima, što upravo govori formula koja je navedena u tabeli 1, pri čemu parametar gama varira u opsegu [0, 1].



Slika 2. Prikaz perfomansi odgovarajućih algoritama

Što se tiče formiranja formule to se radi na sledeći način. Sama formula je suma osnovnih algoritama pomnoženih odgovarajućim koeficijentima u skladu sa treningom algoritma. Takođe posotji i osnovni korektivni faktor a koji se takođe formira na osnovu treninga algoritma u nekoliko iteracija. Konačna formula:

$$f(x) = a + \sum_{i=1}^{N} v \cdot f_i \cdot (Katz(v) + Adar(v) + Jaccard(v))$$

Katz, Adar i Jaccard predstavljaju pojedinačne odbirke koje se množe sa odgovarajućim koeficijentima i koje formiraju konačnu odluku. Takođe oni se množe i sa odgovarajućim v faktorom da bi se uračunala i težina zasnovana na semantičkoj analizi čvorova i interakcije.

4. PLANOVI ZA BUDUĆI RAZVOJ

Jedna od bitnijih stvari koja je predviđena za budući razvoj jeste i predikcija potencijalnih kriticnih novih korisnika, koji bi se kreirali u ovom grafu. Najbitnija stvar koja bi morala da se smisli po tom pitanju jeste kako bi se logicki detektovalo povezivanje novih korisnika, kao i verovatnoća da regularni korisnik pređe u status potencijalnog teroriste. Jedna od mogućih ideja je da se gleda blizina po više faktora sa postojećim teroristima. (blizina po pitanju najkraćeg rastojanja po prijateljima, ali i geografsko rastojanje je bitno uračunati jer je jako velika verovatnoća da će oni biti blizu ili unutar, posebno markiranih kritičnih regiona koji se posebno nadgledaju kao jako rizicni regioni) Naime ovakva struktura rešavanja problema zahteva dosta od administratora same socijalne mreže, ali takve usluge su moguće zarad višeg dobra.

5. ZAKLJUČAK

Navedena tema je jako korisna, po pitanju globalnog društva, naročito u modernom dobu terorizma. Prethodno demnostrirani algoritam donosi neka korisna i zanimljiva rešenja, uz malu cenu privilegovanih informacija. Jedini problemi koji postoje trenutno su kompleksnost podataka i programa koji bi morao da formira i obrađuje informacije. (Naročito ako bi se zahtevala real-time obrada) Takođe tu je i problem privatnosti jer bi takvom programu bile dostupne sve osetljive informacije korisnika u mreži. Takođe moguć je značajan napredak i na polju veštačke inteligencije ako bi se pravio već real-time sistem sa samopodešavanjem.

LITERATURA

[1] Newman, M.E.J. (2001a). Clustering and preferential attachment in grow-ing networks. Physical Review Letters E, 64-67(025102).

[2] Taskar, B., Wong, M.-F., Abbeel, P., & Koller, D. (2003). Link prediction in relational data. In Proceedings of Neural Information Processing Systems (pp. 659–666). Cambridge, MA: MIT Press.

[3] Grossman, J.W. (2002). The evolution of the mathematical research collab-oration graph. Congressus Numerantium, 158, 199–213.

[4] Krebs, V. (2002). Mapping networks of terrorist cells. Connections, 24(3), 41–102.

[5] Kautz, H., Selman, B., & Shah, M. (1997). ReferralWeb: Combining social net-works and collaborative filtering. Communications of the ACM, 40(3), 63–65.