Válaszok Dr. Simon Péter László bírálatára:

1. **Az értekezés jelentős értéke, hogy nagy mennyiségű adat elemzése nyomán hozta létre a szerző elméleti megközelítéssel a modelleket. Ilyen mennyiségű adat elemzése óriási munkát jelent.**

1.1 Mennyire támasztják alá a modelleket a mások által elemzett adatok, az irodalomban található-e erre vonatkozó eredmény?

Ahogy a Bíráló megjegyezte, nagy méretű viselkedési adatok kezelése, tisztítása és elemzése nagy kihívás elé állította a kutatókat, köztük engem is. Ez főleg az adattudományok területének hajnalán volt igaz, amikor még nem voltak elérhetőek az ilyen vizsgálatokra specializálódott programozási és adatkezelési eszközök, így az egész tudományos folyamatot minden esetben a kutatónak kellet kifejlesztenie. Azonban tudományos munkám során a megfigyelt eredmények és az őket leíró modellek általánosíthatóságának és reprodukálhatóságának érdekében minden esetben arra törekedtem, hogy (a) több rendszeren és minél szélesebb paraméterhalmazon erősítsem meg az eredményeket, (b) hogy az adatok elemzése és a modell paraméterezése alaposan legyen dokumentálva, valamint (c) ha az adatvédelmi előírások és a személyes adatok védelme lehetővé tették, a felhasznált adatok egésze, vagy egy reprezentatív mintája a tanulmánnyal együtt publikálva legyen, így teremtve alkalmat másoknak az eredmények újbóli megismétlésére. Ezek a feltételek szükségszerűen kizárták, hogy a megfigyelt valós vagy modellezett jelenségek egy speciális adatbázis, adat kezelési folyamat vagy modell paraméter halmaz eredményeként alakuljon ki, valamint elősegítette az eredmények reprodukálhatóságát is. Ez a megközelítés, mely az adattudomány korai tanulmányai során alakult ki, később a terület standardjává vált, melyet a manapság a folyóiratok is elvárnak a szerzőktől.

Az általam és szerzőtársaim által tett megfigyeléseket és bevezetett modelleket számos esetben másik is megfigyelték és felhasználták, tőlünk függetlenül. Erre bizonyíték a tudományos közleményeimre mutató viszonylag nagy számú hivatkozások száma. Konkrétabban három példát említve:

Karsai, Márton, et al. "Small but slow world: How network topology and burstiness slow down spreading." Physical Review E 83.2 (2011): 025102.

- Egyrészről, ebben a cikkben egy újfajta random referencia modellezési módszert definiálunk temporális hálózatokra, ami később széleskörűen elfogadott és alkalmazott lett a tudományterületen belül. Ennek továbbgondolásaként, pár évvel később egy részletes modell rendszert vezettünk be, ami a temporális hálózatok lehetséges random

referencia modelljeit, mint mikrokanonikus sokaságokat definiálja adott strukturális és temporális kényszerfeltételek mellett [B3_1]. Ugyanakkor, a cikkben megfigyelt jelenség számos további vizsgálatot indukált, amik főleg azt hivatottak eldönteni, hogy a temporális hálózatban megfigyelt bursty kapcsolati dinamikák milyen esetekben lassítják le [B3_2-B3_6] vagy gyorsítják fel [B3_7- B3_9] egy a hálózaton terjedő folyamatot.

Karsai, Márton, et al. "Universal features of correlated bursty behaviour." Scientific reports 2.1 (2012): 397.

- Ez a cikk korrelált bursty vonatok megfigyeléséről és modellezéséről szól. A megfigyelt brusty vonatok méretének hatványfüggvény szerinti eloszlását több más rendszerben is megfigyeltek, mint például [B3_10-B3_13].

Karsai, Márton, Nicola Perra, and Alessandro Vespignani. "Time varying networks and the weakness of strong ties." Scientific reports 4.1 (2014): 4001.

- A cikk fő eredményei azt mutatják meg, hogy társadalmi hálózatok temporális fejlődése során egy személy következő kapcsolatának kiválasztását egy memória folyamat határozza meg mely a már korábban kialakított kapcsolatokat preferálja. Ezt az eredményt mind valós adatokból való megfigyelésekből, mind hálózati modellek segítségével igazoltuk. A cikkben alkalmazott memória folyamatok modelljét több cikkben alkalmaztuk mi és mások is [B3_14-B3_17], valamint a felfedezett memória folyamatok temporális hálózatokra gyakorolt struktúrális hatásait, és egyéb folyamatokra gyakorolt befolyását is további cikkben vizsgálták [B3_18, B3_19].

1.2 Az adatok előkészítése befolyásolhatja-e a modell létrehozását? Például az 1.3.2 szakasz DS2 adatbázisában a 110 millió felhasználó 8 milliárd hívásából számos törlésre került, hogy csak az emberi interakciók kerüljenek be a vizsgálatba, és így egy jelentősen kisebb adatbázis jött létre 80 millió felhasználóval és „csak" egymilliárd összeköttetéssel.

A viselkedési adatok jórészét, így a disszertációban említett adatbázisokat is legtöbbször egy általánosabb viselkedési forma megközelítő megfigyelésére használjuk. Például, mobil kommunikációs adatok a mögöttük meghúzódó társadalmi hálózat közelítő leírására használjuk; a termékek vásárlását társadalmi terjedési folyamatokként értelmezzük; vagy az emberek lokációs adataiból a mobilitásukat vagy társadalmi helyzetüket próbáljuk feltérképezni. Így tekintve az adatokra, azok a rekonstruálandó viselkedési forma szempontjából rengeteg felesleges, vagy félrevezető információt tartalmazhatnak, melyeket a mélyebb tudományos analízis előtt szükségszerű kiszűrni. Sokszor előfordul, hogy az így kapott végleges adat az eredeti adathalmaznak csupán töredéke, de sokkal pontosabb

közelítése az eredeti jelenségnek. Azt, hogy egy tisztított adathalmaz hogyan közelíti az megfigyelendő rendszert úgy tudjuk ellenőrizni, hogy megmérjük, hogy a redukált adat a megfigyelendő jelenségtől független, de a rendszertől elvárt több tulajdonságot teljesíti-e. Egy egyszerű példával élve, ha egy mobilhívási adatbázist társadalmi hálózatok dinamikai fejlődésének a megfigyelésére szeretnénk használni, először ki kell szűrnünk azokat a szereplőket és hívásokat, akik és melyek mögött nem természetes személyek állnak (hanem például cégek, hívásközpontok, vagy maga az adatszolgáltató), és azokat is melyek nem hordoznak valós kommunikációs információt (például nagyon rövid hívások, vagy ön-hívások), valamint minden egyéb, az analízis során megfigyelt adatgyűjtési mellékhatást. Ezek után a visszamaradó adatok validálása az abból konstruált hálózat struktúrális és dinamikai jellemzésével lehetséges. Ekkor azt igyekszünk megmérni, hogy a hálózatban megjelennek a korábban már megfigyelt vagy elvárt hálózati vagy dinamikai tulajdonságok melyek egy társadalmi struktúrát jellemeznek (például heterogén fokeloszlás, magas klaszteresedés, heterogén dinamika, napi aktivitási ciklusok stb.).

Természetesen, ha az eredeti megfigyelésünket a nyers adatokon végeznénk el, akkor más eredményt kapnánk, mint a tisztított adatokon való megfigyelés során. Azonban, habár az adatok tisztítása sokszor nagyban korlátozza a használható adat méretét, ha az adatmanipulációs folyamat során nem egy feltételezett folyamat megfigyelését tartjuk szem előtt, hanem az adattisztítást ettől független feltételrendszer mentén végezzük, az így kapott adatok sokkal bizonyosabb megfigyeléseket biztosítanak, mint maguk a nyers adatok.

2. Milyen lehetőségeket lát a jelölt a most kialakulóban levő, szinte teljesen adatalapú, gépi tanulási módszerekre támaszkodó kutatásokban, amelyekben a sokunk által szeretett modellalkotásnak várhatóan csak periférikus szerep juthat?

Valós természeti vagy társadalmi jelenségek megfigyelésére több modellezési paradigma létezik, melyeket sikerrel alkalmaznak a fizikában és számos más tudományterületen. Azonban ezek a fejlődés nem állt meg, és erre jó bizonyíték a napjainkban népszerű gépi tanulási módszerek, melyeknek fejlődése és alkalmazása igen szembetűnő manapság a tudomány számos területén. Azonban véleményem szerint az egyes modellezési technikák különböző dolgokra alkalmasak, és kevés olyan modellezési rendszer van, mely minden lehetséges oldalról képes lenne leírni egy megfigyelt jelenséget.

Gépi tanulási módszerek nagy előnye, hogy nagy mennyiségű adaton tanítva, elég általános problémamegoldó képességekkel ruházhatjuk fel őket, és így akár az emberinél is pontosabb döntéseket tudnak hozni, illetve ajánlásokat tenni. Ez hatalmas előnyt biztosít olyan esetekben, ha pragmatikusan arra keressük a választ, hogy “hogyan” lehet egy adott kérdésre a legpontosabb választ adni. Ez sok esetben célravezető, ha csupán egy rendszer viselkedését szeretnénk szimulálni annak mélyebb megértése nélkül. Azonban gépi tanulási módszerek bonyolultságukból adódóan leginkább egy fekete dobozként foghatóak fel, így

keveset árulnak el egy rendszer működésének “miértjéről”. Habár napjainkban vannak arra irányuló törekvések, hogy ezeket a rendszereket jobban átláthatóbbá és interpretálhatóbbá tegyék, ez általánosan még nem valósult meg. Miután döntési rendszerek átláthatósága sok területen elengedhetetlen (például orvosi vagy jogi döntéseknél), talán jelenleg ez az egyik legnagyobb kihívás a gépi tanulási módszerek felelőség teljes alkalmazásával kapcsolatban.

Egy fizikai, biológiai, vagy társadalmi rendszerben megfigyelt jelenségek “miértjére” való válaszok keresésére más modellezési módszerek sokkal alkalmasabbak. Egyfelől kauzális statisztikai módszerek vagy random referencia modellek segíthetnek a jelenségek mögött álló oksági struktúra feltérképezésében. Másfelől, mechanisztikus modellek, melyek során egy jelenséget előidéző törvényszerűségeket próbálunk egy modellbe foglalni, sokkal biztosabb válaszokat tudnak adni egy természeti vagy társadalmi jelenség kialakulásának okaira és az ahhoz szükséges körülményekre. Ezen túl, legpontosabban közelítő vagy egzakt analitikus matematikai módszerekkel írható le egy jelenséget, de általában ez csak akkor lehetséges, ha annak csupán leegyszerűsített törvényszerűségeire koncentrálunk. Habár számos más modellezési paradigma létezik és mindegyik megközelítésnek megvannak a maga korlátai, jól kiegészítik egymást és egy széles módszertant biztosítanak a különböző bonyolultságú rendszerek megértéséhez.

Válaszként a Bíráló kérdésére, nem osztom a Bíráló aggodalmát azzal kapcsolatban, hogy a gépi tanulási módszerek kiváltanák a többi paradigma által kidolgozott metodológiákat. Éppen ellenkezőleg, véleményem szerint a gépi tanulási módszerek más egyéb modellezési módszerekkel ötvözve hatalmas potenciált rejtenek magukba. Kezdve egy sokváltozós rendszer paraméterhalmazának optimális feltérképezésétől, hatalmas adathalmazokban keresendő új jelenségek megfigyelésén át, akár új törvényszerűségek felismeréséig, mind jó példák arra ahogy különböző modellezési paradigmák kiegészítve egymást innovatív új módszereket biztosíthatnak a tudomány általános fejlődése érdekében.

Referenciák

[B3_1] Gauvin, Laetitia, et al. "Randomized reference models for temporal networks." *SIAM Review* 64.4 (2022): 763-830.

[B3_2] Vazquez, Alexei, et al. "Impact of non-Poissonian activity patterns on spreading processes." Physical review letters 98.15 (2007): 158702.

[B3_3] Miritello, Giovanna, Esteban Moro, and Rubén Lara. "Dynamical strength of social ties in information spreading." Physical Review E 83.4 (2011): 045102.

[B3_4] Van Mieghem, P., and R. Van de Bovenkamp. "Non-Markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks." Physical review letters 110.10 (2013): 108701.

[B3_5] Jo, Hang-Hyun, et al. "Analytically solvable model of spreading dynamics with non-Poissonian processes." Physical Review X 4.1 (2014): 011041.

[B3_6] Hiraoka, Takayuki, and Hang-Hyun Jo. "Correlated bursts in temporal networks slow down spreading." Scientific reports 8.1 (2018): 15321.

[B3_7] Rocha, Luis EC, Fredrik Liljeros, and Petter Holme. "Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts." PLoS computational biology 7.3 (2011): e1001109.

[B3_8] Takaguchi, Taro, Naoki Masuda, and Petter Holme. "Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics." PloS one 8.7 (2013): e68629.

[B3_9] Rocha, Luis EC, and Vincent D. Blondel. "Bursts of vertex activation and epidemics in evolving networks." *PLoS computational biology* 9.3 (2013): e1002974.

[B3_10] Kikas, Riivo, Marlon Dumas, and Márton Karsai. "Bursty egocentric network evolution in skype." Social Network Analysis and Mining 3 (2013): 1393-1401.

[B3_11] Jo, Hang-Hyun. "Modeling correlated bursts by the bursty-get-burstier mechanism." Physical Review E 96.6 (2017): 062131.

[B3_12] Ba, Cheick Tidiane, Matteo Zignani, and Sabrina Gaito. "Social and rewarding microscopical dynamics in blockchain-based online social networks." Proceedings of the Conference on Information Technology for Social Good. 2021.

[B3_13] Cencetti, Giulia, et al. "Temporal properties of higher-order interactions in social networks." Scientific reports 11.1 (2021): 7028.

[B3_14] Sun, Kaiyuan, Andrea Baronchelli, and Nicola Perra. "Contrasting effects of strong ties on SIR and SIS processes in temporal networks." The European Physical Journal B 88 (2015): 1-8.

[B3_15] Laurent, Guillaume, Jari Saramäki, and Márton Karsai. "From calls to communities: a model for time-varying social networks." The European Physical Journal B 88 (2015): 1-10.

[B3_16] Ubaldi, Enrico, et al. "Asymptotic theory of time-varying social networks with heterogeneous activity and tie allocation." Scientific reports 6.1 (2016): 35724.

[B3_17] Ubaldi, Enrico, et al. "Burstiness and tie activation strategies in time-varying social networks." Scientific reports 7.1 (2017): 46225.

[B3_18] Zino, Lorenzo, Alessandro Rizzo, and Maurizio Porfiri. "Modeling memory effects in activity-driven networks." SIAM Journal on Applied Dynamical Systems 17.4 (2018): 2830-2854.

[B3_19] Kim, Hyewon, Meesoon Ha, and Hawoong Jeong. "Dynamic topologies of activity-driven temporal networks with memory." Physical Review E 97.6 (2018): 062148.

Budapest, 2023. december 4.

Dr. Karsai Márton