Magyar Tudomány 179(2018)7, 991–1000
DOI: 10.1556/2065.179.2018.7.7

ÚJ MÓDSZEREK RÉGI KÉRDÉSEK MEGVÁLASZOLÁSÁRA AZ AKADÉMIAI FELHŐBEN – HÁLÓZATOK ÉS OKSÁGI KAPCSOLATOK FELDERÍTÉSE A TÁRSADALOMTUDOMÁNYOKBAN

NEW METHODS IN THE ACADEMIC CLOUD TO ANSWER OLD QUESTIONS – EXPLORING NETWORKS AND CAUSAL RELATIONS IN SOCIAL SCIENCES

Lévai Péter[1], Telcs András[2]

[1] az MTA rendes tagja, főigazgató, MTA Wigner Fizikai Kutatóközpont
levai.peter@wigner.mta.hu
[2] az MTA doktora, megbízott osztályvezető, MTA Wigner Fizikai Kutatóközpont, MTA–Pannon Egyetem Budapest Rangsor Kutatócsoport, Budapesti Műszaki és Gazdaságtudományi Egyetem Számítástudományi és Információelméleti Tanszék
telcs.andras@wigner.mta.hu

ÖSSZEFOGLALÁS

Cikkünkben röviden bemutatjuk, milyen drámai változáson megy keresztül az informatika fejlődésének hatására a kutatók napi munkája. Részletesen kitérünk annak ismertetésére, milyen, eddig soha nem volt lehetőségek nyílnak meg az asztali és a számítógépes felhőben rendelkezésre álló kapacitások révén a társadalomtudományok (és más területek) kutatói előtt. Példaként röviden ismertetjük, hogyan járulhat hozzá az akadémiai felhő, a nagy kapacitású számítógépek alkalmazása a nemzetközi akadémiai tér, illetve rejtett oksági kapcsolatok feltárásához.

ABSTRACT

In this paper we briefly review the dramatic changes in the researchers' daily practice owing to the development of information technology. We look at it in details how desktop and cloud computational capacities open opportunities never existed before for researchers of social sciences (and practitioners of other fields as well). We present two examples how cloud computing, high performance computing can be applied in the analysis of the global academic space and how hidden causal factors can be revealed.

Kulcsszavak: Moore-törvény, akadémiai felhő, számítógéppel segített kutatás, társadalmi hálók, oksági kapcsolatok

Keywords: Moore law, academic cloud, computer-aided science, social networks, causal relations

BEVEZETÉS

Korunk társadalomtudósai életében egyre gyakoribb, hogy olyan feladatokkal találkoznak, amelyek megoldásához egyrészt a komplex rendszerek és hálózatok területén elért legfrissebb eredmények felhasználása, az ott alkalmazott új algoritmusok alkalmazása szükséges, másrészt jelentős számítási igénnyel lépnek fel, hatalmas számítógépes memóriára és hosszú számolási időre van szükség. Ilyen paradigmaváltáson korábban már átment a fizika, kémia, csillagászat, meteorológia, és ez éri el napjainkban a közgazdaság-tudományt, szociológiát és más társadalom- és humán tudományokat. Ez az új kihívás megkövetelte, hogy újragondoljuk a tudományos szakemberek számítógépes hátterének, igényeinek biztosítását. A Magyar Tudományos Akadémia ezt felismerve teremtette meg a Közép-Európában jelenleg egyedülálló akadémiai felhő szolgáltatást, amely a kutatók számára lehetővé teszi, hogy versenyben maradhassanak a legjobb informatikai hátteret élvező társaikkal.

Cikkünkben bemutatjuk az Akadémiai Felhő nyújtotta lehetőségeket és néhány példát arra, hogyan is lép magasabb szintre egy társadalomtudományi kérdés vizsgálata és megértése az akadémiai felhő segítségével. Konkrét példaként bemutatjuk, hogyan alkalmazhatjuk ezt az újféle számítási diszciplínát az európai egyetemek rangsorolására, a rangsor finomhangolására, idősorok kauzális kapcsolatainak feltárására.

SZÁMÍTÁSI KAPACITÁSOK EGYKORON ÉS NAPJAINKBAN

Nehéz dolgunk van, ha érzékeltetni szeretnénk napjaink számítógépeinek teljesítményét. Az Olvasónak a Moore-törvény juthat először eszébe, miszerint a processzorok kapacitása másfél-kétévente megduplázódik. Hogy ez mit is jelent, azt legegyszerűbb egy személyes példán érzékeltetni. Az 1982-ben megjelent, és egy nyugat-európai magánút során, 1985-ben, leárazás után beszerzett japán CASIO FX-700P programozható számológép már pont elegendő kapacitást nyújtott egykoron, hogy egy fizikus diplomamunka alapszámolásait el lehessen rajta otthon végezni, és ezzel időt nyerjen a tulajdonosa. A CASIO-ban rendelkezésre állt egy 455 kHz-en ketyegő processzor, 2 kB memória és 12 kB tárhely. Egy évtizeddel később, 1992-ben, már egy ötvenszer gyorsabb processzort (25 MHz, azaz 25 000 kHz) és kétezerszer nagyobb memóriát és tárolóhelyet (2 MB memória és 25 MB tároló) tartalmazó személyi számítógép feküdt az asztalon, hogy egy kandidátusi dolgozat kerülhessen ki a billentyűzet alól. Ez a cikk pedig már egy olyan hordozható laptopon íródott, amelyben 2,5 GHz-es sebességgel 6 db processzormag zakatol, 8 gigabyte (GB) memória és 1 terabyte (TB) tárolóegység mellett. A harmincöt év alatt 33 ezerszeresére nőtt számítási kapacitás 2,3 évet ad

ki a Moore-törvényre, de hozzátehetjük, hogy ma már 128 vagy 256 processzormagot tartalmazó chipkártyák is vannak, úgyhogy érvényes a kétéves duplázódás törvénye. Ha pedig a memóriát és tárolókapacitás növekedését nézzük, akkor inkább másfél évenkénti duplázódás tanúi lehetünk. Ne felejtsük el, hogy mindhárom példában új áron 1500 dolláros gépekkel számoltunk.

Ha ilyen gyorsan fejlődött az informatika, akkor felmerül a kérdés, hogy szükséges-e egyáltalán követnünk ezt a teljesítménynövekedést, nem lenne-e elegendő mindaz, amit ma magunkkal hordozunk, vagy akár csak az asztalunkon tartunk? Sőt, talán célszerűbb lenne csak egy kisebb kapacitást magunknál tartani, ami éppen elég a feladataink végrehajtására. Ez ma már meg is valósult, gondoljunk csak az „okos" mobiltelefonjainkra vagy az iPad-ekre, könnyű, kisképernyős laptopjainkra, amelyeket elegendő kétnaponta feltölteni. Ma már egyes okostelefonokban négy processzor ketyeg, s ha akarjuk, akkor a kutatóberendezéseinket, mérőműszereinket is tudjuk távolról irányítani, leolvasni, újraindítani. Ma már nincs technikai akadálya annak, hogy egy-egy új ötletnek szinte azonnal utánajárjunk, egy ismert adatsort új szempontok szerint újraelemezhessünk, akár a buszon ülve is. Egy nehézség azonban jelentkezik: manapság már mindenkinek lehet okostelefonja, iPad-je, sőt laptopja is. Vagyis azon ötletek területén, amelyeket a magunkkal/magunkon hordható infokommunikációs eszközeinkkel megoldhatunk, nagyon nagy a tülekedés: eredményre éhes diákok és kutatók ezrei indulnak ugyanabból a pozícióból. Tulajdonképpen az nyer, akinek hamarabb jut eszébe egy nagyszerű ötlet, és hamarabb tudja ezt hordozható eszközein megoldani, feldolgozni. Úgy is fogalmazhatunk, hogy segédeszközeink, mankóink támogatása mellett megint a szürkeállomány válik fontossá, az ötlet, a koncentrációképesség, a fókuszáltság.

Ugyanakkor létezik egy másik irány is: ha a Moore-törvény továbbra is érvényben marad (és úgy tűnik, hogy igen), akkor a tovább növekedő számítógépes kapacitás lehetővé teszi olyan új kérdések megválaszolását, amelyekre eddig nem is mertünk gondolni, fel sem merült bennünk, mert azt éreztük, hogy ennek még nem jött el az ideje. Azokról az esetekről van szó, amikor az adatállományunk már nem terabyte nagyságú, hanem inkább ezernyi terabyte, s így inkább már petabyte-ban (PB) kell gondolkoznunk. A Big Data Science az az új terület, ahol óriási adatmennyiségeket tudunk átvizsgálni, elemezni, hogy érdekes és új eredményekre juthassunk. Jó példát mutatnak a CERN[1] részecskefizikusai, akik 5000 trillió proton-proton ütközésben begyűjtött 15 000 terabyte-nyi, azaz 15 petabyte adatot rostáltak át ahhoz, hogy háromszáz olyan eseményt találjanak, amely tisztán mutatja a Higgs-bozon megjelenését (és így kísérletileg megalapozhatták a Nobel-díj jogos odaítélését Peter Higgsnek és François Englertnek, akik ötvenöt éve vártak erre a pillanatra). A CERN-ben jelenleg kb. 200 ezer processzormag

[1] CERN: European Organization for Nuclear Research, Európai Nukleáris Kutatási Szervezet.

dolgozik a részecskefizikusoknak, hogy a Higgs-bozonnál is nagyszerűbb, új felfedezések születhessenek. Ehhez még hozzájárul ugyanennyi processzormag a társintézményeknél, ahol szintén adatanalízist végeznek. Ne felejtsük el azonban, hogy ezek a processzormagok továbbra is 2,5–3,5 GHz órajel mellett működnek! Az előrelépés ott történt, hogy a gyártók egyre több magot integrálnak egyetlen chipre (már a 8, 16 vagy 24 magú CPU[2]-k sem ritkák), s ezekre a CPU-kra vonatkozik a Moore-törvény.

Amint láttuk, valójában nem az egyedi processzorsebesség növekszik, hanem a bonyolultsági szint emelkedik. Ami jelentősen megnöveli a megkövetelt szakértelem színvonalát is, hogy ezek az új (és meglehetősen drága) egységek optimálisan kerüljenek kihasználásra. Korábban a kutatók (például a fizikusok) végezték a személyi számítógépeik üzemeltetését, majd rábízták azt a doktoranduszaikra, érdeklődő diákjaikra. De ma már ez nem elegendő. Ma már ismét magasan iskolázott, kiválóan képzett, széles körű tapasztalatokkal rendelkező információtechnológiai szakemberek (IT-mérnökök és IT-technikusok) felügyelik a legkiválóbb rendszereket, biztosítják a felhasználók számára a megígért jelentős kapacitást – és egyúttal versenyelőnyt is biztosítanak.

Korábban az órajelek és adattovábbítási sebességek növelése területén elöl álló gépek képviselték a „High Performance Capacity" (HPC) gépek igencsak szűk csoportját. Sokáig ezek a gépek uralták a TOP-500 rangsor elejét, közepét és végét. Az utóbbi években azonban kiderült, hogy a csúcskategóriás, méregdrága HPC-gépek mellett a nagyon alacsony ár/érték arányt képviselő elemekből is hatalmas, optimálisan kihasználható, úgynevezett „High Throughoutput Capacity" (HTC) egységeket lehet létrehozni. A HTC-egységek megjelenése lehetővé tette, hogy egy adott kérdés megválaszolásánál ne lineárisan, hanem párhuzamosan gondolkodjunk: mit nyerhetünk azzal, ha ezer, vagy akár 20–40 ezer párhuzamosan indítható lépésre bontjuk szét a megoldandó feladatot? Ha időt tudunk nyerni, akkor máris előnyösebb helyzetből indulunk!

De hogyan tegyünk szert több tízezer processzormagra? Hiszen ha gyorsan elvégezzük a feladatunkat, akkor utána nekünk már nem kell az egység. S akkor kapcsoljuk le a gépünket, bocsássuk el a szakembereinket? Mi lesz, ha néhány hónap múlva megint lesz egy szuper ötletünk, és megint szükségünk lesz, mondjuk tízezer processzormagra? Vagy csináljuk azt, hogy üresen üzemeltetjük a rendszert addig, amíg újra szükségünk lesz rá? Hogy tudjuk ezt megfizetni?

A megoldást a felhő (Cloud Computing) jelenti. A felhő maga egy hatalmas számítógépes potenciál több ezer processzormaggal, nem okvetlenül egyetlen helyszínen. A tárterület petabyte nagyságrendű, legtöbbször mágneses szalagra író-olvasó egységgel kiegészítve, hogy az időlegesen nem használt állományok ne foglalják a drága diszkterületet. Meg kell oldani, hogy a felhőt használó *user*ek

[2] CPU: central processing unit, központi feldolgozó egység, a köznyelvben processzor.

úgy érezzék magukat, mintha a „saját" gépükbe léptek volna be, hogy a felhőt bármikor és bárhonnan elérhessék, és ne vesszen el semmi. Ha kevesebb kapacitás is elég, akkor csökkenteni tudják a számukra elkülönített erőforrás nagyságát. Ha pedig néhány napra nagyobb kapacitásra van szükségük, akkor megkaphassák azt – előbb vagy utóbb. Napjainkban így születnek újjá a régi feladványok a felhőkben, így kereshetünk új feladványokra még újabb megoldásokat.

A WIGNER-FELHŐ ÉS AZ AKADÉMIAI FELHŐ

Az MTA Wigner Fizikai Kutatóközpont (FK) néhány éve már szembesült azzal a problémával, hogy a megoldandó feladatok bonyolultsága egyre nőtt, egyre nagyobb számítógépes kapacitásra volt szükség, a kutató kollégáknak egyre több ideje, energiája ment el a technikai háttér biztosítására. Mire sikerült pénzügyi forrást szerezni, majd megvásárolni és beállítani az új számítógépeket, addigra az eredeti feladat sok esetben már el is vesztette fontosságát – jó esetben találtak a kutatók egy új feladatot, amelynél a kutatócsoport által kiépített rendszer felhasználható volt.

Amikor az intézet megépítette a Wigner Adatközpontot, és 2013 januárjában elindult a CERN Kutatóközponttal való együttműködés, lehetőség nyílt arra, hogy modern megoldás szülessen az intézményi szintű problémákra. Először is a Wigner munkatársai a CERN-es kollégákkal együttműködve nagy gyakorlatot szereztek a GRID-hálózatban összekötött számítógépek üzemeltetésében. A CERN számítógépeinek közel fele, mintegy 80 ezer mag és 80 PB tárolókapacitás segíti Budapestről a CERN TIER-0 központi rendszerének működését, szorosan összekapcsolva a genfi székhelyen található 120 ezer maggal.

Először létrejött a különálló Wigner-felhő, amely kb. 1000 processzormaggal és 0,5 PB tárolókapacitással állt a Wigner FK munkatársainak rendelkezésére. A 2015-ös próbaév után, 2016 elejétől a Wigner-felhő már üzemszerűen működött.

Az így szerzett tapasztalatok és a CERN *know-how* alapján a Wigner vezetősége az MTA SZTAKI-val együttműködve meggyőzte az MTA bizottságait és vezetőit, hogy az elszigetelt fejlesztések helyett MTA-szinten egy akadémiai felhő kialakítása lenne a megoldás, hogy a kutatóhálózatban tevékenykedő tudósok megelőzhessék a gazdagabb infrastruktúrával rendelkező országok kutatóit, vagy legalábbis felzárkózhassanak hozzájuk. A SZTAKI-ban és a Wigner Adatközpontban kiépített kapacitás 2016-ban sikeresen teljesítette a próbaidejét, és 2016. októbertől az MTA közösségének rendelkezésére áll a szintén 1000 processzormagból és közel 1 PB tárterületből álló kapacitás. Az akadémiai felhő jelenleg egyedülálló szolgáltatás Közép- és Kelet-Európa kutatási környezetében.

Az akadémiai felhőről részletesebb információk találhatóak az URL1 internetes oldalon. Az oldalon közel negyven projekt összefoglalója is megtalálható, ezzel segítve az érdeklődők kezdeti tájékozódását (URL2).

Az alábbiakban az egyetemi rangsorok elemzése és a kauzális kapcsolatok feltárása szolgál arra példaként, hogy a nagyméretű számítási kapacitás miként teszi lehetővé egy régi probléma újszerű vizsgálatát, új megoldások felfedezését. Ezzel szeretnénk érzékeltetni, hogy a nagy számítógépes kapacitás felhasználása ma már nemcsak a fizikusok, agykutatók és meteorológusok igénye, fontos eszköze, hanem igen hasznos lehet számos különböző társadalomtudományi területen is.

EGYETEMI RANGSOROK

Az egyetemi rangsorok elég hosszú ideje már a közbeszéd, a szakmai vizsgálatok és tudományos kutatás tárgyai. A rangsorok kialakítására irányuló törekvések mögött többek között az az igény húzódik, hogy a társadalom visszajelzést kapjon arról, hogy a felsőoktatási intézmények (továbbiakban: egyetemek) mennyire töltik be a velük szemben támasztott társadalmi elvárásokat, milyen szerepet játszanak a társadalom fejlődésében. Ennek az elvárásnak a tudományos igényű kielégítésére írt ki tematikus kutatócsoportos pályázatot a Magyar Tudományos Akadémia 2017-ben, melynek elnyerése után a Pannon Egyetemen alakult meg a csoport (Budapest Ranking Research Group, BRRG) jelen cikk második szerzője, Telcs András vezetésével.

A csoport munkatársai a hallgatók jelentkezési adatai alapján már korábban is vizsgálták a magyar egyetemek vonzerejét (Telcs et al., 2013). Évente nagyjából százezer diák körülbelül négyszázezer jelentkezési lapot nyújt be, megadva milyen sorrendben preferálja az egyetemeket (karokat, szakokat), amelyekre jelentkezik. A kutatás kiindulásául szolgáló adatbázis tizenöt év rekordjait öleli fel. Ez az adatmennyiség már olyan nagy, hogy egy erősebb személyi számítógép (PC) segítségével is legfeljebb csak egyszerű statisztikákat készíthetünk belőle. A csoport azonban mélyebb elemzéseket is el kívánt végezni, amelyek PC-n megvalósítva soknapos futtatást jelentettek volna. A memória- és teljesítménykorlátokat figyelembe véve, az eredeti szándékhoz képest erősen redukált feladatokat tűztek ki, de még ezek végrehajtása is egy-két napot vett igénybe a rendelkezésre álló számítógépeken.

Ilyen feladat volt a jelentkezések alapján a weblapok rangsorolásánál is alkalmazott PageRank módszer, genetikus algoritmus vagy kimerítő keresés megvalósítása. Ezek, az egyetemek közötti páros összehasonlítást tartalmazó, hálózaton elvégzendő számítások igen erőforrás-, illetve időigényesek. Hasonlóan gépet próbáló feladat a hálózat aggregálása és klaszterezése, a kistérségi preferenciák időbeli és térbeli alakulásának bemutatása (Kosztyán et al., 2015). A vizsgálat közben sok ábra készült, amelyek számítógépes megjelenítése sokat segített a jelenségek megértésében, ugyanakkor további memóriát és kapacitást igényelt hardver oldalról.

Minden eddiginél jobb megoldást kínál a csoport számára az akadémiai felhő és a Wigner-intézet partnerénél, a CERN-ben Jean-Marie Le Goff csoportja által kifejlesztett Collaborativ Spotting (CollSpot, URL3) rendszer. Ezen keretek között tervezzük kialakítani az Academic Space Explorer (ASE) hálózatmegjelenítő és -elemző eszköz modelljét. A BRRG a CollSpotot a rangsorkutatás során három módon is hasznosítani tudja.

Az első, nagyon fontos előnye a nagy mennyiségű adat hálózatos tárolása. A CollSpot a szokásos „Excel-táblázatos", illetve relációs adatbázis helyett az adatokat eleve csúcsok és közöttük futó élek formájában, azaz gráfszemléletben tárolja. A csúcsok és élek attribútumokkal láthatók el. Az így kapott gráf csúcsai objektumosztályokat alkothatnak, amelyek igény szerint klaszterezhetőek, aggregálhatóak, különböző hálózatelemzési eljárásokat lehet elvégezni rajtuk. Az egyetemkutatáshoz ez a gráfszemléletű adattárolás igen szerencsés. Természetes módon adódik, hogy az egyetemeket mint csúcsokat ábrázoljuk, és a szokásos indikátorok lehetnek ezek attribútumai. Ilyen indikátor például a hallgatói létszám vagy az egyetem, illetve a kar kutatói által írt publikációk, a maguk szakterületi besorolásával és a megjelenés adataival (a folyóirat és annak impakt faktora, a megjelenés éve stb.). Már ez a példa is jól mutatja a tárolandó adatmennyiség nagyságát és az adatok között meglévő és megjelenítendő hatalmas és bonyolult kapcsolatrendszert. Mint említettük, a CollSpot lehetővé teszi nemcsak a csúcsok, de a kapcsolatokat reprezentáló élek definiálását is, szintén a maguk attribútumaival. Ez a funkció kiválóan alkalmazható lesz az egyetemi publikációk idézettségi kapcsolatainak a tárolására azok jellemzőivel együtt. Ehhez hasonlóan az ERASMUS-program keretében létrejött diák-, illetve oktatói és dolgozói utazások is jól reprezentálhatóak a hálózat szemléletű adatbázisban.

Mekkora adatmennyiségről is van szó? Ha csak az európai egyetemekre szorítkozunk, az is közel kétezer intézményt jelent. Ezek a vizsgálatba vont tizenöt év alatt évi sokmillió publikációt hoznak létre, azokra tízmilliónyi hivatkozás keletkezik. Hasonló módon évente sok ezer utazás jön létre az ERASMUS keretében. A publikációk, a hivatkozások és az utazások is számos jellemzőjükkel kell hogy tárolásra kerüljenek, hogy csak egyet említsünk: a szakterülettel. Ahhoz, hogy a kutatás során a csoport tagjai az aktuális feladatnak megfelelően gyorsan, egymás zavarása nélkül hozzáférjenek mindezen adatokhoz, megfelelő tároló szerverre van szükség. Ugyanez igaz a mintarendszer működtetésére is. Amennyiben ez az akadémiai felhőben valósul meg, a tárolás és használat költsége töredéke egy saját szerver beszerzési és főleg üzemeltetési költségének.

A hálózatos adattárolás mellett a kutatók egy, az elemzést támogató eszközt is kapnak a CollSpot révén. A rendszer a tárolt hálózatot képes megjeleníteni, azon kívánság szerint műveleteket végrehajtani. A CollSpot számos megjelenítést készen ajánl fel, és továbbiakat is be tud fogadni. A hálózatmegjelenítés egyes opciói menüből érhetők el, másokat magán a megjelenített hálózaton navigálva

lehet kezdeményezni (szép demonstráció nézhető meg a Collaboration Spotting honlapján, URL4).

A hálózatot transzformálhatjuk, aggregálhatjuk, klaszterezhetjük, egyes részeit kiemelhetjük, és áttekinthetjük időbeli fejlődését. Természetesen a rendszer a kiválasztott objektumokat a kért attribútum szerint sorba is rendezi és megjeleníti, ami már az egyetemi rangsorok létrehozásához visz közelebb bennünket.

Végezetül, a kidolgozott új elemzési, megjelenítési és rangsorolási eredmények a rendszerbe beépíthetőek, és remekül mutathatók be. Hogy csak egy példát említsünk az utóbbira: az egyetemek közötti idézettségi kapcsolatokból kiindulva általános vagy akár szakterületspecifikus, a tudományos teljesítmény szerinti rangsor készíthető a PageRank algoritmus, illetve annak megfelelő általánosítása segítségével.

A CallSpot tehát a nagy mennyiségű adat hálózatos formában való tárolása mellett a hálózat elemzését és bemutatását is elősegíti, mindehhez azonban, tekintettel az adatmennyiségre és az adatok közötti kapcsolatok komplexitására, erős számítógépes kapacitásra van szükség. A CallSpot kényelmes és kutatóbarát szolgáltatásai olyan, erősen számításigényes feladatokat eredményeznek, amelyek az akadémiai felhőben lényegesen hatékonyabban végezhetőek el, mint saját eszközön.

KAUZÁLIS KAPCSOLATOK FELTÁRÁSA

Az oksági kapcsolat a tudomány szent grálja. Világunk megértése oksági kapcsolatok feltárásán keresztül történik, és történelmünk kezdetéig nyúlik vissza. Valóban a papok imádsága hozza el a Nílus áradását biztosító életadó esőt, a nap és a csillagok állása határozza meg, mekkora lesz az áradás? Egy jelenség vizsgálata során alapvető a kérdés: mi okozza? Hasonlóan gyakori kérdés, hogy a vizsgált jelenség egyik, illetve másik komponense az ok, illetve az okozó-e. Arisztotelész négy oksági kapcsolatot határozott meg. Francis Bacon kettéválasztja a gondolkodást fizikára és metafizikára, lehetővé téve a természettudományos oksági kapcsolat definiálását és vizsgálatát. A modern természet- és társadalomtudomány az ilyen jellegű kérdéseket már igyekszik minél racionálisabban, azaz adatokra alapozva megválaszolni. Norbert Wiener majd Clive Granger (vö. URL5) vezették be a statisztikus kauzalitás fogalmát, amely két alapelven nyugszik:

1. az ok megelőzi az okozatot, és
2. az ok figyelembevételével az okozat jobban jósolható, mint nélküle.

Granger maga az idősorok ARIMA-modelljére dolgozta ki a nevét viselő kauzalitás vizsgálatát. Ezt számos általánosítás, illetve alternatív módszer kidolgozása követte, amelyeket a nemlineáris, entrópiaalapú elemzések széles családjába lehet besorolni.

A kauzalitásról beszélve mindenképpen kiemelendőnek tartjuk a Floris Takens időeltolásos beágyazási elméletén (Takens, 1981) alapuló, George Sugihara által kidolgozott „converging cross mapping" módszert (Sugihara et al., 2012). Sajnos, abban az esetben, amikor a Granger- vagy Sugihara-módszer által kimutatott oksági kapcsolat áll fenn, e módszerek segítségével nem lehet kizárni azt, hogy az A és B jelenségeknek létezik egy rejtett közös oka is.

Somogyvári Zoltán és Telcs András csoportjukkal olyan új módszert dolgoztak ki, amely képes a rejtett ok meglétét kimutatni annak közvetlen megfigyelése, megfigyelhetősége nélkül. A módszer kiindulópontja Takens elmélete. Ennek lényege, hogy a megfigyelt idősorokból külön-külön és együttes megfigyelésükből, időeltolásos beágyazással egy-egy geometriai alakzat, sokaság készíthető. Ennek dimenziója jellemzi az idősor információtartalmát. Ha A, B a két megfigyelt idősor, jelölje dimenziójukat d(A), d(B) és az együttesükét d(A,B). Az alábbi táblázat dimenziók közötti kapcsolatok és a kauzális összefüggések ekvivalenciáit tartalmazza.

dimenziók kapcsolata	kauzális kapcsolat
$d(A) < d(B) = d(A,B)$	A okozza B-t
$d(B) < d(A) = d(A,B)$	B okozza A-t
$d(A) = d(B) = d(A,B)$	A és B kölcsönosen okozzák egymást
$d(A),d(B) < d(A,B) = d(A)+d(B)$	A és B független
$d(A),d(B) < d(A,B) < d(A)+d(B)$	A-nak és B-nek közös oka van

A módszer mintavételi adatokra épül, ezért inherensen statisztikus természetű, a belőle levonható következtetés is az. Ennek megfelelően a módszer a fenti öt esethez bayesi gondolatmenettel valószínűségeket rendel.

Eltekintve az adatok előfeldolgozásától a módszer alkalmazásához egy legalább ötdimenzós paramétertérben kell a jó paramétereket meghatározni. Ez csak a módszer egyenkénti újrafuttatásával lehetséges, ami nyilvánvalóan sokszoros párhuzamos futtatással valósítható meg ideálisan. A módszer alkalmazására álljon itt egy egyszerű példa.

Tekintsük a New York-i tőzsdén forgalmazott részvények kellően hosszú idősorait. Döntsük el melyek a „fontos" papírok, amelyek mozgása „okozza" a többi árváltozását. A feladat igen hasonló a sokváltozós regresszióelemzésből ismert modellválasztási feladathoz. A feladat megoldása során egyes változókra vagy változócsoportokra újra és újra le kell futtatni az elemzést. A feladattól függően n idősor esetén ez a szám n-től akár 2^n-ig változhat. Nyilvánvaló, hogy az utóbbi értékhez közeli esetekben a számítás párhuzamosítás nélkül elképzelhetetlen.

Párhuzamosítással, az akadémiai felhőben a feladat tíz-tizenöt értékpapírra még észszerű idő alatt elvégezhető.

Példáink rávilágítanak arra, hogy új tudományos eredmények eléréséhez nemcsak a természet- és élettudományok területén, hanem a társadalomtudományok területén is szükséges a nagy számítógépes kapacitás és tárolás. Az akadémiai felhő nyújtotta lehetőségek kiaknázása lényegesen javíthatja a nemzetközileg is versenyképes eredmények létrehozását. Az akadémiai felhő, valamint a Wigner Kutatóközpont és annak Komputációs Tudományok Osztálya nyitott a társadalomtudományok művelői felé. Célunk az, hogy igényeikhez, kutatási problémáikhoz igazodó szakmai és erőforrás-támogatást adjunk, ezzel is hozzájárulva kutatásuk sikeréhez.

IRODALOM

Kosztyán Zs. T. – Telcs A. – Török Á. (2015): Felsőoktatásba jelentkezők preferenciáinak térbeli és időbeli szerkezete, teljesítményfüggése. *Statisztikai Szemle*, 93, 10. 917–942. http://www.ksh.hu/statszemle_archive/2015/2015_10/2015_10_917.pdf

Telcs A. – Kosztyán Zs. – Török Á. (2013): Hallgatói preferencia-sorrendek készítése az egyetemi jelentkezések alapján. *Közgazdasági szemle*, LX, március, 290–317. http://www.kszemle.hu/tartalom/letoltes.php?id=1371

Sugihara, G. – May, R. – Ye, H. et al. (2012): Detecting Causality in Complex Ecosystems. *Science*, 338, 6106, 496–500. DOI: 10.1126/science.1227079, https://www.researchgate.net/publication/230895543_Detecting_Causality_in_Complex_Ecosystems

Takens, F. (1981): Detecting Strange Attractors in Turbulence. *Lecture Notes in Mathematics*, 898, 1, 366–381. http://www.crcv.ucf.edu/gauss/info/Takens.pdf

URL1: https://cloud.mta.hu/

URL2: https://cloud.mta.hu/projektek

URL3: http://collspotting.web.cern.ch/

URL4: http://collspotting.web.cern.ch/sites/collspotting.web.cern.ch/files/HTML/maps/technogram.html#technogram.json

URL5: http://www.scholarpedia.org/article/Granger_causality