



**„Gazdaságinformatika  
Doktori Iskola”**

## **TÉZISGYŰJTEMÉNY**

**Balogh Zoltán**

**Személyes adatok gyűjtése és profilírozás az interneten**

című Ph.D. értekezéséhez

**Témavezető:**

**Dr. Racskó Péter, CSc**

egyetemi docens

Budapest, 2017

**Számítástudományi Tanszék**

**TÉZISGYŰJTEMÉNY**

**Balogh Zoltán**

**Online környezetben begyűjthető adatok jellemzői és azokból kinyerhető  
személyes jellemzők: egy Budapesti Corvinus Egyetem e-learning  
rendszerében történt kutatás bemutatása**

című Ph.D. értekezéséhez

**Témavezető:**

**Dr. Racskó Péter, CSc**

egyetemi docens

© „Balogh Zoltán”

## Tartalomjegyzék

<b>I.</b>	<b>KUTATÁSI ELŐZMÉNYEK ÉS A TÉMA INDOKLÁSA.....</b>	<b>5</b>
I.1.	A KUTATÁSOK BEMUTATÁSA.....	6
<b>II.</b>	<b>A FELHASZNÁLT MÓDSZEREK .....</b>	<b>8</b>
<b>III.</b>	<b>AZ ÉRTEKEZÉS EREDMÉNYEI .....</b>	<b>10</b>
III.1.	AZ ESZKÖZBÖNGÉSZŐKRŐL BEGYŰJTHETŐ ADATOK JELLEMZŐI .....	10
III.2.	AZ EGYETEMI POLGÁROK PSZICHOLÓGIAI JELLEMZŐI ADATVÉDELMI SZEMPONTBÓL .....	11
III.3.	SZEMÉLYES INFORMÁCIÓ KINYERÉSE WEBES ADATOKBÓL.....	12
<b>IV.</b>	<b>FŐBB HIVATKOZÁSOK .....</b>	<b>13</b>
<b>V.</b>	<b>A TÉMAKÖRREL KAPCSOLATOS SAJÁT PUBLIKÁCIÓK JEGYZÉKE.....</b>	<b>17</b>

## I. Kutatási előzmények és a téma indoklása

Ahogy az internet a hétköznapjaink részévé vált, a weboldalakat megjelenítő alkalmazások – a böngészők – fejlődése is felgyorsult. Ez lehetővé tette a weboldalak számára, hogy egyre több adatot érjenek el a látogatóikról. Az online tartalmak megtekintésével megismerhetővé váltak a látogatók fogyasztóinak preferenciái és böngészési szokásai. Ezen adatok elemzésével a látogató személyre szabott tartalmat kaphat és ezzel egyidejűleg felfedhetővé válnak a személyes jellemzői is. A személyre szabott tartalmak következtében megjelent a „filter bubbles” jelensége, amely során a felhasználó az észlelt tulajdonságaihoz illő tartalmakhoz fér hozzá, viszont nincs kontrollja afelett, hogy mi az, amit fogyaszthat és mi az, amit nem. A felhasználók profilozása miatt a kezdeti anonim web ma már nem az.

A technológiai fejlődés következtében megismerhetővé vált a weboldalak látogatóinak preferenciái, aminek következtében megjelentek az online ajánlórendszerek. Például, az amerikai online műsorszóró vállalat, a Netflix ajánlórendszere hozzávetőlegesen az esetek 80%-ban befolyásolja a látogatókat a tartalmak fogyasztásában. (Carlos & Neil, 2015) Emiatt nem meglepő, hogy a cég igencsak érdekelt az általuk használt algoritmusok hatékonyságának növelésében. Emiatt 2006 és 2009 között 1 millió dolláros Netflix Díjjal ösztönözték vállalkozó szellemű vállalkozókat az ajánlórendszer hatékonyságának javítására. A győztes csapat által kidolgozott algoritmus hatékonyságában 10,06%-kal múlta felül a Netflix által használtat. (Lohr, 2009) Az internetes szolgáltatók komoly erőfeszítéseket tesznek ügyfeleik profiljának minél pontosabb meghatározására, mert ez számukra versenyelőnyt, sokszor pedig a pályán maradás feltételét jelenti.

Az információszerzés és felhasználás területén jelenleg paradigmaváltás figyelhető meg: amíg korábban a felhasználók pusztán a korszerű keresőmotorok képességeit kihasználva találták meg a számukra releváns információt, mára a keresőmotor használatán felül az online szolgáltatások a látogatók preferenciájának megismerésével képesek eldönteni, hogy egy adott tartalom a látogató számára érdekes lehet-e, azaz a tartalom jellemzőit és a látogató preferenciáit összevető algoritmus dönti el, hogy az a látogató számára érdekes-e. A preferenciákat összeállító, profilozó algoritmust használó oldalak (keresőoldalak, hírportálok és közösségi oldalak) használatával a látogatók implicit módon hozzájárulnak a detektálható jellemzőik gyűjtéséhez, majd ezen adatok alapján az oldal tartalmakat ajánl számukra fogyasztásra. A weboldalak által használt ajánlórendszerekről a látogatók sok esetben nem tudnak és használatuk mellőzésére az esetek többségében nincs mód.

## I.1. A kutatások bemutatása

Kutatásaim során arra kerestem a választ, hogy a weboldalak számára mely hozzáférhető adatokból lehetséges a látogató tulajdonságaira következtetni, mely adatok alkalmasak a profilok összeállítására és ez hogyan történik. Kutatásaimat az alábbiak szerint építem fel:

Az első kutatási cél a nem professzionális, **egy domain alól elérhető weboldalak<sup>1</sup> vizsgálata**. Exploratív kutatás keretében az internetezéshez használt böngészőkből, valamint az internetezéshez használt hardverről és annak szoftver-környezetéből kinyerhető adatokat elemezve következtettem a felhasználó személyes jellemzőire. Az elemzési fázisban a felhasználók és személyes paramétereik között csak triviális kapcsolatot sikerült kimutatni, a magukat szándékosan felfedni nem kívánó látogatók esetében nem sikerült személyes jellemzőket megállapítani. Megvizsgáltam a böngészők számára **hozzáférhető paraméterek bizonytalanság-csökkentő erejét** is. Felmértem az egy domain alól elérhető weboldalak és közösségi oldalak látogatóikról elérhető adatok mennyiségét és minőségét. A látogatókról összegyűjtött adatok bizonytalanság-csökkentő képessége megmutatja, hogy a mintán belül mekkora valószínűséggel található meg egy egyed, és összehasonlíthatóvá válik a kinyert paraméterek különböző csoportosításainak információhordozó ereje.

A második kutatási cél a Budapesti Corvinus Egyetem **közösségi oldalak** által összegyűjtött, **kinyilvánított preferenciákból** kinyert **személyes tulajdonságok** alapján a látogatók csoportosítása, majd a kapott csoportok összevetése a myPersonality Project (Stillwell & Kosinski, 2012) során összegyűjtött adatokból készített klaszterekkel. Azt mutatom be, hogy a felhasználók által önként szolgáltatott adatokból hogyan lehet személyiségre vonatkozó következtetéseket levonni. Természetesen az adatokat felhasználás előtt anonimizáltam. A kutatás során a Facebook-tól letöltött egyénekhez köthető „Like” adatbázist elemeztem.

A közösségi hálózatok aktív tagjai az online tartalmakról alkotott tetszésüket a „Like”<sup>2</sup> gombra történő kattintással is kifejezhetik. Ez a látogatókhoz köthető információ az közösségi

---

<sup>1</sup> Általában az egy domain alatt elérhető weboldalak képesek hozzáférni a böngészéshez használt eszköz, a rajta lévő operációs rendszer és a böngésző valamennyi tulajdonságához (céges weboldak, hírportálok, blogok, webáruházak), feltétel, hogy nincs az oldalnak más weboldalakra beépülő adatgyűjtő modulja, amellyel a látogatók preferenciáit vagy böngészési jellemzőit lehetséges feltérképezni, a látogatók nem regisztrálják magukat az oldalra, amely esetben a beazonosítás triviálissá válna

<sup>2</sup> Facebook Like gomb (2010 második negyedév): a felhasználók kifejezhetik a tetszésüket egy weben található tartalom iránt. Ezzel a lépéssel azokról a weboldalakról is képes a Facebook adatokat gyűjteni a felhasználóiról a közösségi oldal meglátogatása nélkül

hálózatokon elérhető. A kutatás során a Budapesti Corvinus Egyetem polgárainak „Facebook Like”-jait elemeztem pszichológiai API segítségével (Kielczewski, 2017), majd az egyéneket a kapott személyes jellemzőik alapján nem felügyelt tanulási módszerekkel klasztereztem. A kutatás szintén feltáró jellegű, a Budapesti Corvinus Egyetem polgárainak és a myPersonality Project résztvevőinek a pszichológiai API által visszaadott személyes tulajdonságokból képzett látogatói klaszterek közötti különbségeket mutatja be.

A harmadik kutatási célom annak kikísérletezése volt, hogy hogyan lehet **a látogatók nem személyes jellemzőiből következtetni személyes jellemzőikre**. Az Apriori algoritmus (Gautam, Ghodasara, & Parsania, 2014) használatával a Budapesti Corvinus Egyetem polgárainak e-learning környezetbeli viselkedését elemezve arra kerestem a választ, hogy a látogatók mely személyes tulajdonságaikra lehetséges online viselkedésükből következtetni.

A látogatók személyes tulajdonságuk alapján történő csoportokba rendezése üzletileg jól hasznosítható eredményt hoz, ui. az egyes csoportoknak célzott reklámok küldhetőek. A kezdetek óta a Facebook egyik üzleti stratégiai alappillére a közösségi hálózatokban rejlő hirdetési felület adta lehetőség kiaknázása. (Jeffrey, 2012) A weboldalak látogatói az üzlet szempontjából meghatározó tulajdonságaik alapján csoportosíthatóak.

## II. A felhasznált módszerek

A kutatásaimhoz szükséges minta begyűjtéséhez saját adatgyűjtő alkalmazást fejlesztettem, mivel az interneten találhatóak elégtették ki a megfelelő minőségű minta iránti alábbi kívánalmaimat:

- a hozzáférhető adatok legszélesebb körét képes legyen elmenteni
- alacsony szintű hozzáférést biztosítson az összegyűjtött adatokhoz
- az ismert blokkoló alkalmazások ne legyenek képesek megakadályozni a működését
- adatgyűjtő alkalmazás működése közben az átlagos felhasználó látogató ne tudja, hogy adatait kezelem

Az adatgyűjtő alkalmazás kliensoldali részét Javascriptben, a szerveroldali részét PHP-ben fejlesztettem le, az adatokat MySQL adatbázisba mentettem el. Az adatgyűjtést követően az adatok előfeldolgozását Pentaho-val végeztem el. (A földrajzi pozíció település és utcanévre történő átváltása Google Maps API használatával, böngésző pontos típusának meghatározása a HTTPUserAgent mezőből, meglátogatott tantárgyak oldalbetöltésekhez rendelése stb.)

Az Egyetemi polgárok pszichológiai jellemzőinek vizsgálata során a Budapesti Corvinus Egyetem polgárainak a Facebook Like-jaikból (kinyilvánított preferenciák) kinyert személyes tulajdonságait elemeztem. A kutatás keretében megismerhető, hogy a felhasználók által önként szolgáltatott adatokból miként lehetséges személyes jellemzőket kinyerni, megismerhető az Egyetemi polgárok személyes adatainak megosztási hajlandósága, valamint bemutatja a hozzáférhető személyes adatok sokaságát. Az egyéneket a személyes tulajdonságaik alapján klaszterekbe rendeztem, majd az eredményeimet összehasonlítottam a myPersonality Project kutatásának adatgyűjtési fázisának eredményeképpen létrejött mintából képzett klaszterekkel.

A University of Cambridge Psychometrics Center kutatói az online profilozás mérföldkövét tették le 2012-ben, amikor 58 000 olyan önkéntes Facebook felhasználóval pszichológiai tesztekkel töltettek ki, majd ezt követően elemezték a látogatók Facebook-on található személyes adataikat is. Az elemzések eredményeinek egybevetését követően fejlesztettek egy mindenki számára elérhető API-t, amely a Facebook-os like-ok alapján képes meghatározni a vizsgálat alanyának pszichológiai jellemzőit. (Kosinski, Stillwell, & Graepel, 2013)



A Facebook Like-okból a személyes tulajdonságokat az „Apply Magic Sauce” pszichológiai API-val nyertem ki és azokat látogatókhoz társítva adatbázisba mentettem el. A elemzéséhez SPSS statisztikai programcsomagot használtam, az alapvető statisztikai számítások elvégzésén felül K-közép és hierarchikus klaszterező eljárásokat alkalmaztam.

Az Egyetemi polgárok pszichológiai jellemzői adatvédelmi szempontból című kutatásom során az előzőekben ismertetett Facebook-os adatokból kinyert személyes jellemzőket használtam fel és az online környezetben tanúsított viselkedésük és a környezeti jellemzőkből következtettem a látogatók személyes tulajdonságaira. A kutatásban vizsgált mintában a személyhez nem köthető változók az attribútum halmaz elemei, a személyes tulajdonságok pedig a potenciális osztályváltozók. Az elemzés során több klasszifikációs algoritmust, többek között a népszerű Apriori algoritmust is használtam, amely nagy adathalmazokban képes hatékonyan asszociációs szabályok keresésére. (Agrawal & Srikant, 1994) Az Apriori algoritmust több adatbányászati alkalmazásba is implementálták, én a Java alapú Wekát használtam. (Witten, Frank, & Hall, 2011)

Az Apriori algoritmus gyakori elemhalmazok elemzésével nyeri ki az asszociációs szabályokat. Az algoritmus kihasználja a gyakori elemhalmazok részhalmaza is gyakori axiómát, ezáltal az algoritmus hatékonyan tudja csökkenteni a kinyert szabályok számát, melyek közül csak azok lesznek érdekesek, amelyek egy minimális support (támogatottság) értéknél nem kevesebb. (Agrawal & Srikant, 1994) Már egészen kis minták esetében is rengeteg különböző asszociációs szabály nyerhető ki, az algoritmus csak azokat tartja meg, amely viszonylag nagy számú példányra érvényes.

### III. Az értekezés eredményei

Kutatásaim során egyaránt vizsgáltam a nem professzionális, egy domain alól elérhető weboldalak és professzionális a közösségi oldalak vagy hirdetési ügynökségek által hozzáférhető adatokat. A kutatásaimból és a feldolgozott irodalomból levont konzekvenciákat a böngészés kiemelt aktorai számára az alábbiakban összegzem:

- **Weboldalak látogató egyének:** az egy domain alól elérhető weboldalak számára viszonylag kevés információ érhető el, ebből korlátozottan, de lehetséges személyes jellemzők kinyerésére, a kinyilvánított preferenciák miatt a közösségi oldalak és a hirdetési ügynökségek által hozzáfért adatokból lehetséges a személyes jellemzőkre következtetni. Minél pontosabban ismert a látogató személyes jellemzői, annál jobban a látogatóhoz illeszkedő tartalmakat és reklámokat lehetséges küldeni számára, ami a „filter bubbles” jelenségéhez vezet. Ezt elkerülendő célszerű tudatosan használni a közösségi oldalakat, keresőmotorokat és minden olyan weboldalt, amelybe a hirdetési ügynökségek reklámhordozókat ágyaztak be.
- **Adatvédelmi hatóság:** az Európai Unió kiberbiztonsággal foglalkozó felelős szervezete a „European Union Agency for Network and Information Security” (ENISA), amelyet 2004-ben hoztak létre annak érdekében, hogy ajánlásokat tegyen, valamint a politikai szempontok kialakításánál és bevezetése során kulcsszerepet vállaljon az Európai Unió számára. (ENISA, 2017)
- **Szoftverfejlesztő/CIO:** A felhasználói/látogatói viselkedés és preferenciákból történő azonosítás megelőzése miatt érdemes az alkalmazások felkészítése a tárolt információk fokozott védelmére. Fontos az ENISA által kiadott „privacy by design” elv és a rá épülő keretrendszerek (ISO/IEC 29100) alkalmazása, amelynek célja az alkalmazások azok tervezési fázisában az információ védelmére való felkészítése. (Danezis, és mtsai., 2015)

#### III.1. Az eszközböngészőkről begyűjthető adatok jellemzői

A feltáró jellegű kutatásom célja annak bemutatása, hogy az egy domain alól elérhető weboldalak számára a látogatók által használt böngészők és eszközökről elérhető adatok használhatóak-e a látogató azonosítására és követésére. A kutatásomban elemeztem, hogy a látogató eszközböngészőjéből hozzáférhető adatok milyen mértékben járulnak hozzá ehhez.

Egy oldalletöltés alkalmával akár néhány száz kilobyte-nyi adat hozzáférhető a meglátogatott webhely számára, amelyből adatbányászati módszerekkel értékes – akár személyhez köthető – információk nyerhetőek ki, valamint ezen adatok segítségével lehetséges az egyének beazonosítása és követése, személyes adataik felhasználása nélkül. A rendelkezésre álló paramétereket a következő négy csoportba rendeztem:

- a böngésző környezetének jellemzői
- a szoftver és hardver környezet jellemzői
- a hardver környezet és a böngésző jellemzőit és
- az egy átlagos munkamenet ideje alatt állandó paramétereket.

A kutatás eredményeképpen azt tapasztaltam, hogy minél több változót használok a felhasználó azonosításához, annál nagyobb vizsgálatba bevont változók együttes bizonytalanság eloszlató képessége. Ez azt jelenti, hogy egy ismert sokaság elemei közül nagy bizonyossággal be tudjuk azonosítani az egyén által használt eszközböngészőt, feltéve, hogy a munkamenet ideje alatt állandó paraméterei ismertek. Ebből az is következik, hogy amennyiben meg szeretnénk nehezíteni a követőink dolgát, célszerű proxy-n keresztül minél kevesebb paramétert megosztani a követő weboldalak és harmadik felek számára.

### III.2. Az egyetemi polgárok pszichológiai jellemzői adatvédelmi szempontból

A myPersonality Project keretében összegyűjtött nemzetközi adatokban és a saját adatgyűjtés eredményeképpen létrejött adatokban klasztereket képeztem nem felügyelt tanulási algoritmusokkal, majd a kapott eredményeket összevettem. A kutatás eredményeképpen megállapítható, hogy a különböző forrású adatokban hasonló egyének csoportjai vettek részt, annak ellenére, hogy a myPersonality Project keretében gyűjtött adatok főként az angolszász országok polgárai, a saját kutatásomban pedig a Budapesti Corvinus Egyetem polgárai vettek részt.

A kutatásom idejében a Facebook like-ok pszichológiai elemzése során kinyerhető az egyénre jellemző Big5 személyiségi jellemzők, az étellel való elégedettség, intelligencia, kor, nem, szexuális orientáció, érdeklődési kör, politikai beállítottság, hitvallás és családi állapot. A pszichológiai API a Facebook like-okból visszaadott változók bizonyosságát a Pearson-féle korrelációs együtthatóval jellemzi, amely szerint többségük (kor, nem, szexuális orientáció,

érdeklődési kör, politikai beállítottság, hitvallás és családi állapot) igen erős lineáris kapcsolatot, míg néhány (Big5, étellel való elégedettség és intelligencia) paraméter közepesen erős kapcsolatot mutat. A különböző forrású mintákon K-közép és hierarchikus klaszterképző eljárásokat futtattam, amelynek eredményeképpen létrejött klaszterek mindkét módszer esetében ugyanazokat eredményezte, így azok stabilnak mondhatók.

A vizsgált mintákban (myPersonality Project és a Corvinus Egyetem polgárai) kettő nagyon hasonló klaszter lelhető fel: a magas nyitottság, neurotikusság és extrovertáltság értékkel rendelkező tartalommegosztók, akik nagy valószínűséggel nincsenek kapcsolatban és a komoly kapcsolatban lévő, kompromisszumra képes, lelkiismeretesek egyének. Emellett a myPersonality Project mintájából kimutatható még a közösségi hálózaton kevés időt töltő egyének csoportja, akik feltételezhetően a Corvinus Egyetem polgárainak mintájában is jelen vannak.

### III.3. Személyes információ kinyerése webes adatokból

A Budapesti Corvinus Egyetem polgáraitól gyűjtött mintában kimutatható összefüggés található a böngésző személyisége és az általa használt szoftver és hardver környezet tulajdonságai, valamint a látogató online viselkedése között.

Gyakori elemhalmazok részhalmazait keresttem az adatbányász körökben népszerű Apriori algoritmussal, amelyek közül a magas konfidenciaszinttel és támogatottsággal rendelkezőket vizsgáltam. (Agrawal & Srikant, 1994) Az algoritmus talált érdekes szabályokat, amelyekkel a látogatók 10%-a esetében magas megbízhatósággal képes az intelligenciájukat és étellel való elégedettségüket előre jelezni.

A talált szabályok csak a vizsgált mintára érvényesek, de megfelelően képesek a webes adatok adatbányászati algoritmusokkal történő potenciál érzékeltetésére.

## IV. Főbb hivatkozások

- Abramson, M., & Aha, D. W. (2013). User authentication from Web browsing behavior. *Florida Artificial Intelligence Research Society Conference* (old.: 6). St. Pete Beach: AAAI Press.
- Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases* (old.: 487-499). San Francisco: Morgan Kaufmann Publishers Inc.
- Andreas, P., & Marit, H. (2010. augusztus 10). *Privacy and Data Security, TU Dresden, Faculty of Computer Science*. Forrás: A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management: [http://dud.inf.tu-dresden.de/Anon\\_Terminology.shtml](http://dud.inf.tu-dresden.de/Anon_Terminology.shtml)
- Barabási, A. (2010). *Villanások - a jövő kiszámítható*. Budapest: Helikon Kiadó Kft.
- Bodon, F. (2010. február 28). *Adatbányászati algoritmusok*. Budapest, Magyarország.
- Chris, H. J., Ashkan, S., Nathaniel, G., & Dietrich, W. J. (2012. január 1). Behavioral Advertising: The Offer You Can't Refuse. *Harvard Law & Policy Review* vol. 6, old.: 273-296.
- Clarke, R. (1999). Internet Privacy Concerns Confirm the Case for Intervention. *Communications of ACM*, 60-67.
- Cser, L., & Fajszki, B. (2004). *Üzleti tudás az adatok mélyén - Adatbányászat alkalmazói szemmel*. Budapest: Budapesti Műszaki és Gazdálkodástudományi Egyetem.
- Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.-H., Le Métayer, D., Tirtea, R., & Schiffner, S. (2015. január 12). *European Union Agency for Network and Information Security*. Letöltés dátuma: 2017. május 14, forrás: Privacy and Data Protection by Design: <https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design>
- Davenport, D. (2002. április). Anonymity on the Internet: Why the Price May Be Too High. *Communications of the ACM* vol. 45, no. 4, old.: 33-35.
- Domokos, M. N. (2013). Az EU új adatvédelmi szabályozása – avagy „keep bangin' on the wall of Fortress Europe”. *Jogi Fórum*, 1-46.
- Eckersley, P. (2013, január 26). *Electronic Frontier Foundation - Defending your rights in the digital world*. Retrieved april 19, 2013, from A Primer on Information Theory and

Privacy: <https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy>

- ENISA. (2017). *European Union Agency for Network and Information Security*. Letöltés dátuma: 2017. május 14, forrás: About ENISA: <https://www.enisa.europa.eu/about-enisa>
- Escobido, M., & Gillian, S. (2013). Can Personality Type be Predicted by Social Media Network Structures? *The Asian Conference on Psychology & the Behavioral Sciences*. Osaka: The International Academic Forum.
- Európai Bizottság. (2015. július 11). *A személyes adatok védelme*. Forrás: Európai Bizottság honlapja: [http://ec.europa.eu/justice/data-protection/index\\_hu.htm](http://ec.europa.eu/justice/data-protection/index_hu.htm)
- France, B., & Robert, C. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. *MISQ, volume 35, issue 4*, 1017-1041.
- Haig, Z., Kovács, L., Ványa, L., & Vass, S. (2014). *Elektronikai hadviselés*. Budapest: Nemzeti Közszerológiai Egyetem.
- Hunyadi, L., & Vita, L. (2006). *Statisztika közgazdászoknak*. Budapest: Központi Statisztikai Hivatal.
- Jia-Ching, Y., Chu-Yu, C., & Vincent, T. S. (2012). Mining web navigation patterns with dynamic thresholds for navigation prediction. *IEEE Computer Society 2012* (old.: 614-619). Hangzhou: IEEE.
- John, L., Manuel, B., & Luis, A. v. (2004. február). Telling Humans and Computers Apart Automatically. *Communications of the ACM*, 57-60. Letöltés dátuma: 2013. július 14, forrás: [http://www.cs.cmu.edu/~biglou/captcha\\_cacm.pdf](http://www.cs.cmu.edu/~biglou/captcha_cacm.pdf)
- Kang, R., Brown, S., & Kiesler, S. (2013). Why do people seek anonymity on the internet?: informing policy and design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2657-26666.
- Kennedy, H. (2006). Beyond anonymity, or future directions for internet identity research. *New Media & Society, Vol 8, Issue 6*, 859-876.
- Kiss, A. (2015. február 23). *Az adatokhoz, adatbázisokhoz kapcsolódó jogi szabályozás I.* (A. Kiss, Előadó) Budapest.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 5802-5805.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. In U. o. Kenneth Wachter (Szerk.),

- Proceedings of the National Academy of Sciences of the United States of America*.  
110, old.: 5802–5805. Berkeley: PNAS.
- Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2013. october 19).  
Manifestations of user personality in website choice and behaviour on online social  
networks. *Machine Learning, June 2014, Volume 95*, old.: 357-380.
- Kosinski, M., Las Casas, D., Paulo Pesce, J., Quercia, D., Stillwell, D., Almeida, V., &  
Crowcroft, J. (2012). Facebook and Privacy: The Balancing Act of Personality,  
Gender, and Relationship Currency. *Sixth International AAAI Conference on Weblogs  
and Social Media*. Dublin: ICWSM.
- Kovács, E. (2014). *Többváltozós adatelemzés*. Budapest: Typotex.
- Nan, Z., Aaron, P., & Haining, W. (dátum nélk.). *An Efficient User Verification System via  
Mouse*.
- Nemeslaki, A., Kis, G., Duma, L., & Szántai, T. (2004). *e-Business: Üzleti modellek*.  
Budapest: ADECOM Kommunikációs Szolgáltató Rt.
- Peter, O., David, G., David, L., Warren, F., & Jonathan, N. B. (2005). Continuous Identity  
Verification. *Jur*, 20-24.
- Racsó, P. (2012). A számítási felhő az Európai Unió Egén. *Vezetéstudomány*, old.: 1-16.
- Shababi, C., Zarkesh, M. A., Adibi, J., & Shah, V. (1997). Knowledge discovery from users  
web page navigation. *26th IEEE International Conference on research in Data  
Engineering*, (old.: 20-29).
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System  
Technical Journal*, 379-423.
- Stillwell, D. J., & Kosinski, M. (2012). *myPersonality project: Example of successful  
utilization of online social networks for large-scale social research*. Cambridge,  
University of Cambridge, UK: The Psychometrics Centre.
- Stillwell, D., Kosinski, M., Rust, J., & Wang, N. (2012. february 3). Can Well-Being be  
Measured Using Facebook Status Updates? Validation of Facebook's Gross National  
Happiness Index. *Social Indicators Research vol 115, issue 1*, old.: 483-491.
- Szabó, A. (2010). *Random Forests - Véletlen erdők*. Letöltés dátuma: 2017. január 8, forrás:  
Adatbányászat és Keresés Csoport:  
<https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2011/randomforests.pdf>
- Személyes adatok feldolgozása vonatkozásában az egyének védelméről és az ilyen adatok  
szabad áramlásáról, 95/46/EK (Az Európai Parlament és a Tanács 1995. október 24).

- Voulodimos , A. S., & Patrikakis , C. Z. (2009. december). Quantifying privacy in terms of entropy for context aware services. *Identity in the Information Society*, 2(2), 155-169.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining - 3rd edition*. Burlington: Morgan Kaufmann.
- Youyou, W., Kosinski, M., & Stillwell, D. (2015. január 27). Computer-based personality judgments are more accurate than those made by humans. *PNAS*, old.: 1036-1040.



## V. A témakörrel kapcsolatos saját publikációk jegyzéke

### Tudományos könyvek, könyvrészek

Balogh Zoltán [2012]: Internetes anonimitás aktorai. In *Gazdaság, Társadalom II. A jövő és 2012, Arisztotelész*, ISBN: 987-963-87721-8-3, pp. 6-16.

### Magyar nyelvű referált szakmai folyóiratcikkek

Racsó Péter, Szommer Károly, Balogh Zoltán [2014]: *Az online világban hagyott virtuális lábnyomokban rejlő információ és azok veszélyei*. In *Vezetéstudomány, volume XLV., issue 2014. 7-8., BCE*, ISSN: 0133-0179, pp. 97-104.

Szommer Károly, Balogh Zoltán [2016]: *Geotagging használata Magyarországon*. In *Minőség és megbízhatóság, EOQ MNB Egyesület*, ISSN: 0580-4485, pp. 140-147

### Angol nyelvű referált szakmai folyóiratcikkek

Balogh Zoltán [2012]: Identification in eLearning Environment. In *SEFBIS Journal 2013 No.8*, ISSN: 1788-2265, pp. 81-86

Rétallér Orsolya, Balogh Zoltán [2015]: *Specialities of Psychological Traits of Citizens of Corvinus University of Budapest*. In *Hadmérnök X. Évfolyam 4. szám*, ISSN 1788-1919, pp. 193-204

### Egyéb (angol)

Balogh Zoltán [2012]: *Anonymity over the internet*. Proceeding of *Cyter conference 2012*, ISBN: 978-80-01-05072-9, pp. 1

Balogh Zoltán [2012]: *Anonymity over the internet*. Proceeding of *18th International ICE-Conference on Engineering, Technology and Innovation*, ISBN: 978-1-4673-2275-1, pp. 377-383

Balogh Zoltán [2012]: *Do-not-track*. Proceeding of *Professzorok az Európai Magyarországiért 2012*, ISBN: 978-963-88433-7-1, pp. 9

Balogh Zoltán [2012]: *Potential dangers of using the web*. Proceeding of *GIKOF e-Journal*, pp. 8

Szommer Károly, Balogh Zoltán [2015]: *Dangers of sharing platforms used by people of different personalities*. Proceeding of *ACOMP 2015, IEEE Computer Society*, ISBN-13: 978-1-4673-8234-2, pp. 7-11.

Balogh Zoltán [2016]: *Mining web data with Apriori algorithm*. Proceeding of *SKIMA 2016*, ISBN: 978-1-5090-3298-3, pp. 10