

INFORMAČNÍ BULLETIN



České statistické společnosti

Ročník 24, číslo 2, červen 2013

ZPRACOVÁNÍ HROMADNÝCH DAT DISTRIBUCE VELIKOSTI AEROSOLOVÝCH ČÁSTIC V REÁLNÉM ČASE A JEJICH PREZENTACE V EVROPSKÝCH I CELOSVĚTOVÝCH DATABÁZÍCH

MEASUREMENT AND ANALYSIS OF NEAR REAL TIME DATA ON PARTICLE SIZE DISTRIBUTION IN ATMOSPHERIC AEROSOL

Zdeněk Wagner¹, Miroslav Fridrich²,
Jaroslav Schwarz³, Vladimír Ždímal³

Adresa: ¹Termodynamická laboratoř E. Hály, Ústav chemických procesů
AV ČR, v. v. i., Rozvojová 135, 165 02 Praha 6
wagner@icpf.cas.cz

²Oddělení IT, Ústav chemických procesů
AV ČR, v. v. i., Rozvojová 135, 165 02 Praha 6
fridrich@icpf.cas.cz

³Oddělení aerosolových a laserových studií, Ústav chemických procesů
AV ČR, v. v. i., Rozvojová 135, 165 02 Praha 6
schwarz@icpf.cas.cz, zdimalv@icpf.cas.cz

Poděkování: Práce byla podporována grantem GAČR č. 209/11/1342.

Abstract: The objective of this work is to measure particle size distribution of an atmospheric aerosol using scanning mobility particle sizer (SMPS) that takes aerosol samples at Košetice Observatory. The measured data are sent in near real time (NRT) to our institute for further processing and for providing the results to a central database. Data available centrally can serve theoreticians to develop and test regional and global meteorological and climatic models. Data provided in near real time can be used for improving the quality of weather nowcasting and for risk assessment when unusual events such as volcanic eruptions or spills of harmful substances from industrial plants occur. Calculation and online presentation of charts on the web enable us early detection of unusual phenomena as well as diagnostics of instrument faults that helps to speed-up service and thence to reduce instrument downtime.

Keywords: Atmospheric aerosol, time series, distribution function, non-parametric estimate, mathematical gnostics.

1. Úvod

Jednou z nejvýznamnějších složek znečištění ovzduší jsou částice atmosférického aerosolu. Mají několik klíčových funkcí. Jako kondenzační jádra jsou zárodky kapek v mracích a zásadně ovlivňují dynamiku koloběhu vody v přírodě. Zároveň přímo i nepřímo ovlivňují celkovou energetickou bilanci Země, neboť tvoří překážku tokům záření z meziplanetárního prostoru směrem k Zemi i ve směru opačném. Silně nelineární interakce aerosolových částic s elektromagnetickým zářením, popsaná Mieho teorií, má také za následek extrémní vliv aerosolových částic takzvaného akumulárního modu, tedy o rozměru řádově stovek nanometrů, na dohlednost v troposféře. V neposlední řadě se čím dál více prokazuje vliv atmosférického aerosolu na zdraví lidí, zejména té složky jemné aerosolové frakce, která vznikla spalovacími procesy.

Ve všech výše uvedených případech je rozhodujícím parametrem velikost aerosolových částic. Proto jsou zejména v poslední dekádě vyvíjeny nové detektory aerosolových částic, schopné měřit rozdělení velikosti částic ve vysokém velikostním rozlišení. Nejčastěji se jedná o takzvané skenovací třídiče velikosti částic, zkráceně SMPS, které jsou schopny vzorkovat aerosol s několikaminutovou frekvencí a poskytují velikostní rozlišení až 64 kanálů na řád přes takřka celou frakci jemných částic mezi jednotkami nanometrů a jedním mikrometrem. Tyto aerosolové spektrometry SMPS jsou umísťovány do monitorovacích sítí, jako je například síť evropských aerosolových stanic EUSAAR [1], která poskytuje v téměř reálném čase (near real time) údaje o dynamice velikostních rozdělení přízemního atmosférického aerosolu do hlavních evropských aerosolových databází.

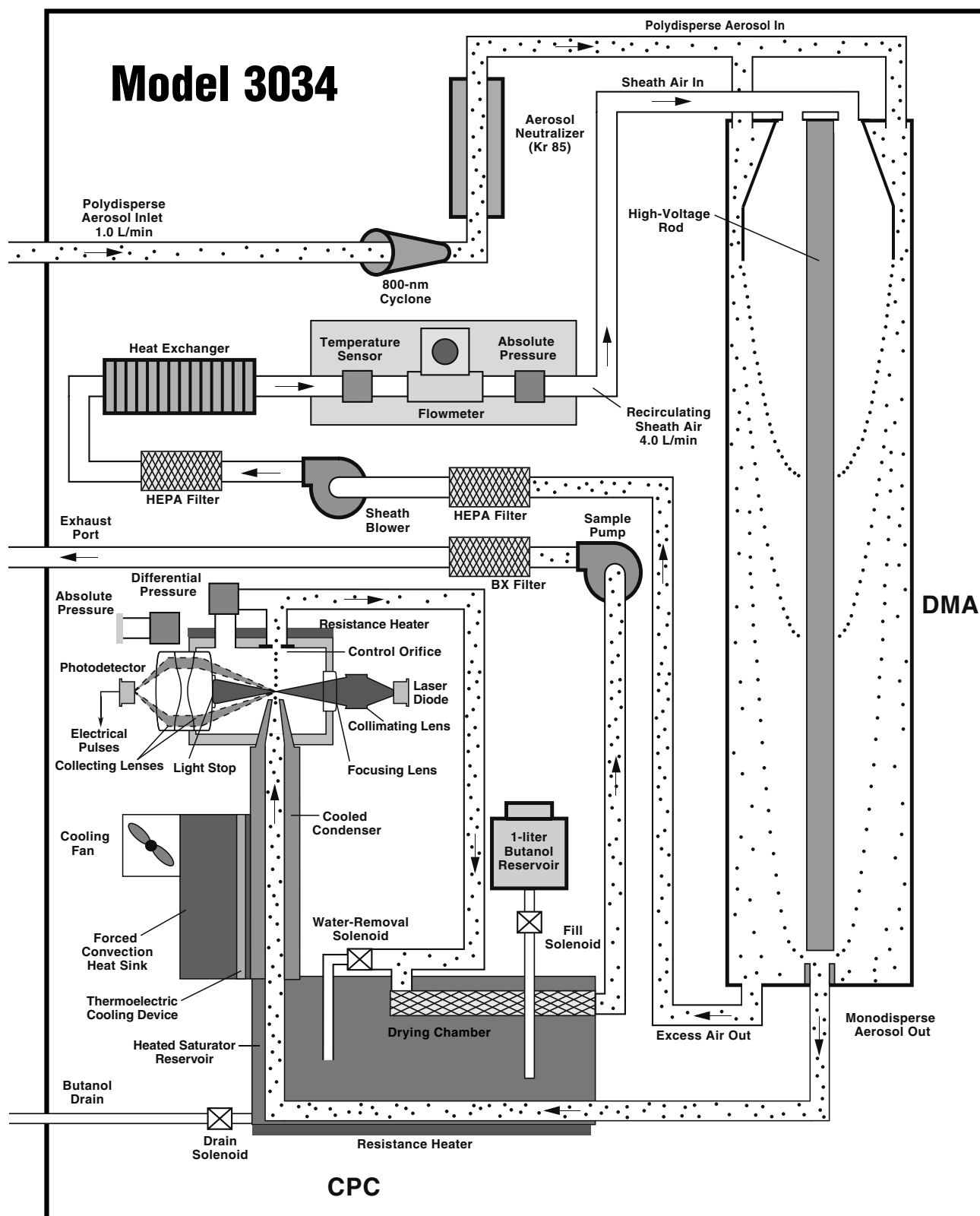
V předložené práci jsme se zabývali vývojem metod efektivního zpracování časových řad získaných spektrometrem SMPS a jejich prezentací v databázích.

2. Měření a přenos dat

Vzorky atmosférického aerosolu jsou odebírány na observatoři Českého hydrometeorologického ústavu v Košeticích. Tato observatoř se nachází na Českomoravské vrchovině v zemědělské krajině mimo souvislé osídlení a mimo dosah přímých zdrojů znečištění. Nadmořská výška je 534 m a zeměpisné souřadnice jsou N 49°35', E 15°05'. Měření probíhá nepřetržitě od 1. 5. 2008.

Měření je prováděno spektrometrem SMPS (Scanning Mobility Particle Sizer), jehož schéma, převzaté z literatury [2], je ukázáno na obrázku 1.

Spektrometr SMPS odebírá vzorky v ekvidistantních časových intervalech 5 min. Třídič (DMA) rozdělí vzorek podle průměru částic na 71 veli-



Obrázek 1: Schéma spektrometru SMPS

kostních tříd jejichž geometrické středy se pohybují přibližně od 10 nm do 1000 nm. K dělení dochází tak, že aerosolové částice jsou nabity a následně unášeny proudem vzduchu v elektrickém poli. Částice jsou elektrickým polem vychýleny kolmo ke směru proudu vzduchu. Vychýlení závisí na průtoku vzduchu v třídiči, napětí elektrického pole a hmotnosti částice. Částice má obvykle jeden elementární náboj, proto lze z velikosti vychýlení, napětí, průtoku a hustoty vypočítat její velikost. Větší částice mohou nést více elementárních nábojů, pravděpodobnostní rozdělení počtu nábojů v závislosti na velikosti částice je známo.

Počet částic v každé velikostní třídě je následně změřen kondenzačním čítačem (CPC). Účinnost čítání je závislá na velikosti částice a určuje se kalibrací.

Po naměření každého spektra jsou do souboru v řídicím počítači zapsány nekorigované počty částic a hodnoty teploty, tlaku, průtoků plynů a jejich relativních vlhkostí. Software dodávaný s přístrojem provede ještě korekci na vícenásobné elektrické náboje a do souboru zapíše i korigovaná data. Korekci na účinnost čítače software neprovádí. Tuto korekci je nutno zahrnout do následného zpracování.

Naměřená data se přenášejí veřejnou datovou sítí do počítače v Ústavu chemických procesů. Zde se konvertují do formátu NASA/Ames 1001, jehož specifikaci lze nalézt na internetové stránce projektu [3]. Soubor je pak odeslán do databází EUSAAR, EMEP a GAW-WDCA na server v Norsk institutt for luftforskning, Kjeller, Norsko.

3. Zpracování a prezentace dat

Data získaná na observatoři Košetice slouží nejen pro účely projektu EUSAAR, ale i pro hodnocení sezónní variability a pro srovnávání dat naměřených na jiných lokalitách. Data ve formě histogramu nejsou pro další zpracování vhodná. Výhodnější je vyjádření pomocí spojitě distribuční funkce. Aerosolové částice pocházejí z více zdrojů, distribuce je tedy zpravidla multimodální. K jejímu stanovení lze použít buď parametrický, nebo neparametrický odhad. Protože měření je automatické a nepřetržité, musí být metoda zpracování dostatečně rychlá, aby ji bylo možno použít v reálném čase, a současně spolehlivě fungující bez nutnosti ručních zásahů. Tyto požadavky jsou tedy výchozím bodem pro volbu metody zpracování.

Základní problém parametrického odhadu spočívá v tom, že děje v atmosféře jsou velmi komplikované. Je obtížné najít univerzální, obecně použitelný model. Obvykle se používá směs dvou až tří lognormálních distribucí.

Vychází se přitom ze zkušenosti, že tento model často dobře popisuje reálnou distribuci velikosti aerosolových částic.

Neparametrické odhady narážejí na jiné úskalí. Základem je určení správné šířky jádra (bandwidth). Při nesprávné hodnotě hrozí underfitting nebo overfitting. Použití statistických neparametrických odhadů tedy může vyžadovat jisté doladění parametrů [4, 5]. Bayesovský přístup je elegantní, ale obvykle výpočetně náročný [6]. Použitelnost pro výpočty v reálném čase je tedy značně omezená. Výhodným se tedy jeví využití estimační lokální distribuční funkce odvozené z matematické gnostiky [7]. Algoritmus pro její stanovení je dostatečně rychlý a metoda poskytuje objektivní kritérium výpočtu parametru měřítka, které bylo odvozeno na základě teoretických úvah.

Stanovení distribuční funkce není jedinou úlohou. Jak bude ukázáno dále, naměřené hodnoty mohou být poškozeny různými vlivy. Jednou příčinou jsou krátkodobé poruchy přístroje, druhou příčinou je dočasná přítomnost blízkého zdroje aerosolových částic. Poškození způsobené krátkodobou poruchou je nežádoucí v každém případě. Rozpoznání blízkého zdroje částic je však často užitečné. Chceme-li určit zdravotní zátěž pro lidský organismus, jsou data zahrnující lokální zdroje důležitá. Pokud naopak data slouží k vývoji a testování fyzikálních modelů atmosféry, je žádoucí tyto vlivy odfiltrovat. Prvním krokem zpracování je proto filtrace naměřených koncentrací částic v jednotlivých velikostních třídách v časové doméně.

Oba přístupy ke stanovení distribuční funkce, semiparametrický a neparametrický, byly porovnány v předchozí práci [8]. Přestože výsledky obou přístupů jsou v mnoha ohledech srovnatelné, je parametrická metoda v principu schopna popsat pouze distribuci s nejvýše třemi mody. Proto jsou dále používány pouze algoritmy vycházející z matematické gnostiky [7].

3.1. Filtrace a interpolace v časové doméně

Hlavní účel filtrace koncentrace aerosolových částic v časové doméně byl již zmíněn. Filtrace však má i další důvod. Meteorologické veličiny mohou být měřeny v jiných časových intervalech než distribuce velikosti aerosolových částic. Bylo by tedy vhodné, kdybychom uměli v časové řadě interpolovat. V předchozí práci [9] byla vyvinuta metoda založená na robustní regresi řadou Čebyševových polynomů 2. druhu v posuvném okně. Hladkost výsledné filtrované řady je závislá na volbě stupně polynomu a šířce posuvného okna.

Šířka posuvného okna má také značný vliv na zpoždění filtru. Jestliže provádíme filtraci tak, že je okno symetrické a hledáme filtrovanou hodnotu ve středu okna, pak zpoždění filtru je dáno polovinou šířky okna. Taková filtrace je pak nepoužitelná pro zpracování dat v reálném čase, protože velké

zpoždění je v mnoha aplikacích nepřijatelné. Filtraci je však možno provést v nesymetrickém okně.

Protože rychlost výpočtu je nejdůležitějším faktorem, nepoužíváme při filtraci přímo časové údaje, ale malá celá čísla. Měření byla provedena v ekvidistantních časových intervalech $t_1, t_2, \dots, t_K$. Průměry částic v jednotlivých velikostních třídách označíme $d_1, d_2, \dots, d_R$. Nechť počty částic v daných velikostních třídách naměřené v čase t_k jsou $n_{1,k}, n_{2,k}, \dots, n_{R,k}$. Nechť u, v jsou malá přirozená čísla a $T = \max(u, v) + 1$. Abychom určili filtrovanou hodnotu v čase t_k , zvolíme posuvné okno $t_i, i = k - u, k - u + 1, \dots, k + v$ a definujeme pomocnou proměnnou:

$$\kappa_{k,i} = \frac{i - k}{T}. \quad (1)$$

Časovou závislost počtu částic ve velikostní třídě s poloměrem d_r v posuvném okně v okolí t_k pak vyjádříme tímto vztahem:

$$\phi_{r,k}(\kappa_{k,i}) = \sum_{m=0}^M c_{r,k,m} U_m(\kappa_{k,i}), i = k - u, k - u + 1, \dots, k + v \quad (2)$$

Parametry $c_{r,k,m}$ získáme robustní regresi podél influenční křivky [10] odvozené z matematické gnostiky a s využitím *minimum penalty estimation* [11, 12] pro zlepšení poměru signál/šum. Při regresi použijeme apriorní váhu:

$$w_i = \sqrt{1 - \kappa_{k,i}^2} \quad (3)$$

Hodnota w_i hraje roli *zapomínací váhy*, ale současně zvyšuje stabilitu výpočtu, neboť Čebyševovy polynomy 2. druhu jsou na intervalu $[-1, 1]$ s touto vahou ortogonální.

Rovnice (2) je použitelná pro filtraci v časové doméně, ale nelze ji přímo využít k interpolaci, neboť v obecném případě $\phi_{r,k}(t) \neq \phi_{r,k+1}(t)$ pro $t_k < t < t_{k+1}$. Použijeme tedy interpolační vzorec

$$n_r(t) = \lambda(\tau) \phi_{r,k}\left(\frac{\tau}{T}\right) + [1 - \lambda(\tau)] \phi_{r,k+1}\left(\frac{\tau - 1}{T}\right) \quad (4)$$

kde

$$\tau = \frac{t - t_k}{t_{k+1} - t_k}, \quad \lambda(\tau) = \frac{1 + \cos \pi \tau}{2}. \quad (5)$$

Lze snadno dokázat, že funkce (4) je spojitá, spojitě diferencovatelná.

Při vyhodnocování rovnice (2) není nutno počítat hodnoty jednotlivých Čebyševových polynomů, ale využívá se rekurentní Clenshawův vzorec, který je numericky stabilnější.

3.2. Stanovení distribuční funkce

Distribuci velikosti částic v atmosférickém aerosolu popisujeme neparametricky estimační lokální distribuční funkcí odvozenou z matematické gnostiky [7]. Přesný postup stanovení byl popsán v předcházející práci [8]. Je však vhodné zdůraznit, že distribuce je heteroskedastická, což souvisí s tím, že aerosol obsahuje částice z různých zdrojů a s různým stářím.

3.3. Prezentace dat

Vypočtené hodnoty jsou ukládány tak, aby numerické hodnoty mohly být použity k dalším analýzám. Pro rychlou orientaci je však vhodné grafické zobrazení. Grafy musí být dostupné i vzdáleně. Byla proto zvolena prezentace na webové stránce <http://hroch486.icpf.cas.cz/Košetice/>, přičemž grafy jsou generovány a ukládány na webovou stránku bezprostředně po výpočtu. Jsou tedy dostupné téměř v reálném čase s ohledem na to, že k přenosu naměřených dat z košetické observatoře dochází v hodinových intervalech a další zpoždění je způsobeno filtrací.

4. Výsledky

Klíčovým problémem při filtraci v časové doméně je volba šířky okna a stupně polynomu. Na rozdíl od algoritmu stanovení gnostické distribuční funkce, kde teorie poskytuje objektivní kritérium, zde takové kritérium není k dispozici. Obecně lze říci, že stupeň polynomu musí být nejméně 3, aby rovnice (2) byla schopna popsat inflexní bod. Použití apriorní váhy (3) zajišťuje numerickou stabilitu regrese, nicméně použití polynomů příliš vysokého stupně vede k rozkmitání především v bezprostředním okolí rychlých změn koncentrace a zejména interpolace pak může poskytovat zkreslené výsledky. Vhodné parametry je nutno vyzkoušet filtrací časové řady ve zvoleném krátkém časovém úseku. Ukazuje se, že variabilita časového vývoje koncentrace aerosolových částic je malá, takže parametry, použité pro filtraci dat měřených na Krétě v rámci projektu SUB-AERO [13], dávají stejně dobré výsledky i při filtraci dat z observatoře v Košeticích.

Je-li cílem filtrace zachování rychlých změn způsobených obvykle lokálními zdroji, chceme tuto informaci získat s minimálním zpožděním. V rovnici [2] tedy použijeme parametry $M = 5$, $u = 9$, $v = 1$. Snížení zpoždění filtru použitím hodnoty $v = 0$ zhoršuje kvalitu filtrace při rychlých změnách koncentrace, vyšší hodnota $v = 2$ poskytuje horší kvalitu filtrace než hodnota $v = 1$, a to kvůli riziku rozkmitání. K rozkmitání dochází též při použití vyššího stupně polynomu než 5.

Časová řada s odstraněnými rychlými změnami způsobenými lokálními zdroji je obvykle využívána *ex post* k dalším srovnávacím analýzám. Kvalita filtrace je tudíž prvořadá a zpoždění filtru není na závadu. Osvědčily se proto hodnoty $M = 3$, $u = v = 19$.

Z distribuční funkce se následně určují polohy módů a koncentrace částic, jak celkové, tak ve zvolených velikostních intervalech. Velikostní intervaly jsou voleny tak, aby koncentrace co nejlépe vypovídaly o dějích probíhajících v atmosféře.

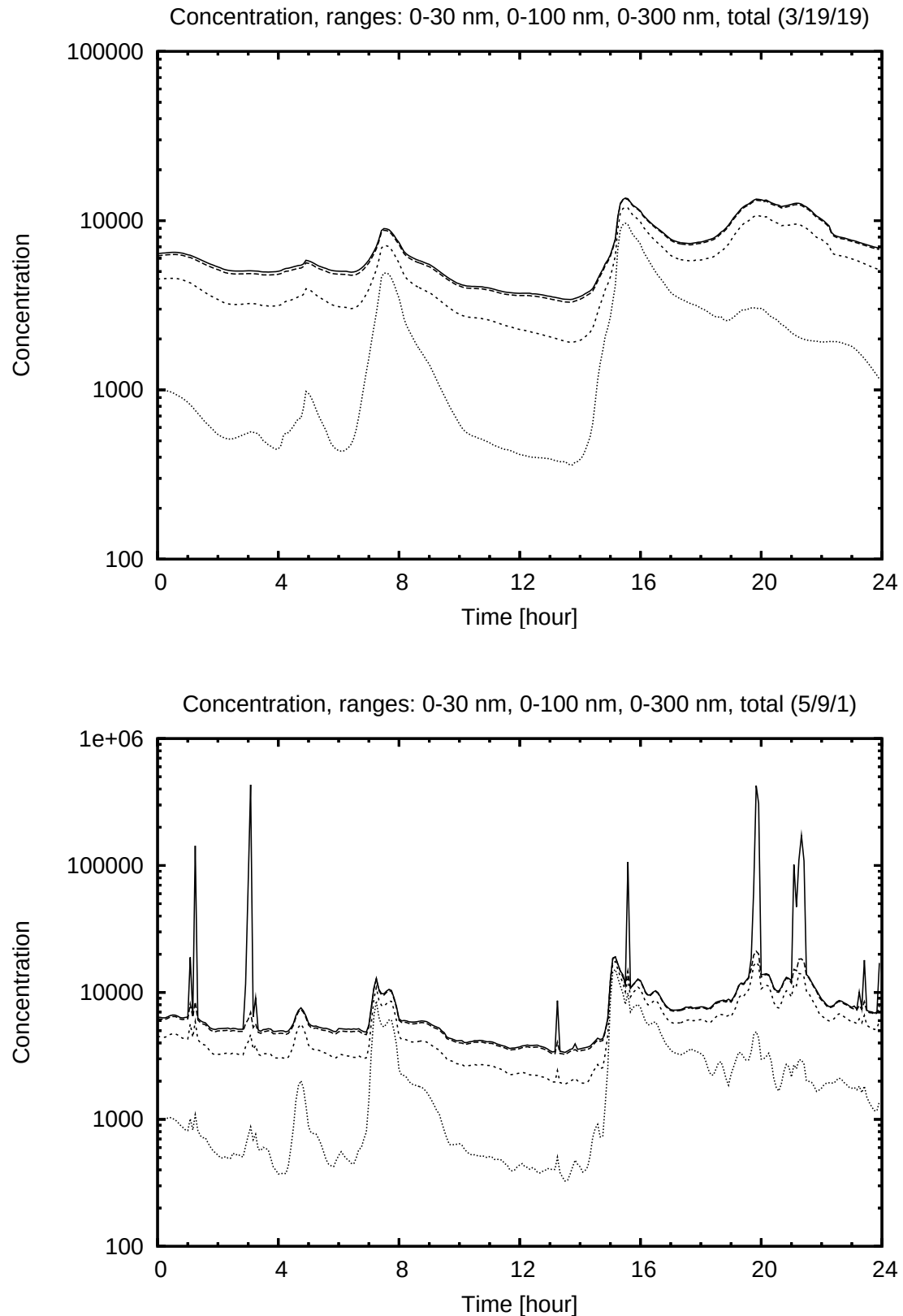
Obrázek 2 ukazuje časovou závislost celkové koncentrace částic a koncentrace ve třech velikostních intervalech v době, kdy vlivem poruchy docházelo k výbojům na elektrodě. K zapálení výboje je nutné vysoké napětí, proto měření koncentrace malých částic není ovlivněno, ale při měření koncentrace velkých částic dochází k chybě dosahující téměř dvou řádů. Výboj se objevil vždy v několika po sobě jdoucích měřeních. Matematicky se tedy jevil jako rychlý děj a při filtraci s parametry $M = 5$, $u = 9$, $v = 1$ byl zachován. Filtrace s parametry $M = 3$, $u = v = 19$ poškození dat vyhladila. Zejména srovnání obou výsledků pomáhá k rychlému zjištění, že přístroj neměří správná data.

Další obrázky ukazují situaci v době, kdy se vyskytoval lokální zdroj částic, v tomto případě se jednalo o sečení trávy v areálu observatoře. Sečení probíhalo téměř celý den, proto je zvýšení koncentrace patrné na výsledcích z obou metod filtrace. Největší rozdíl v koncentracích na obrázku 3 je patrný zejména v době, kdy byla tráva sečena v bezprostřední blízkosti vzorkovacího místa. Z obrázku 4 je zřejmé, že filtrace s parametry $M = 3$, $u = v = 19$ i v tomto případě výrazně vyhladila časovou závislost polohy módů.

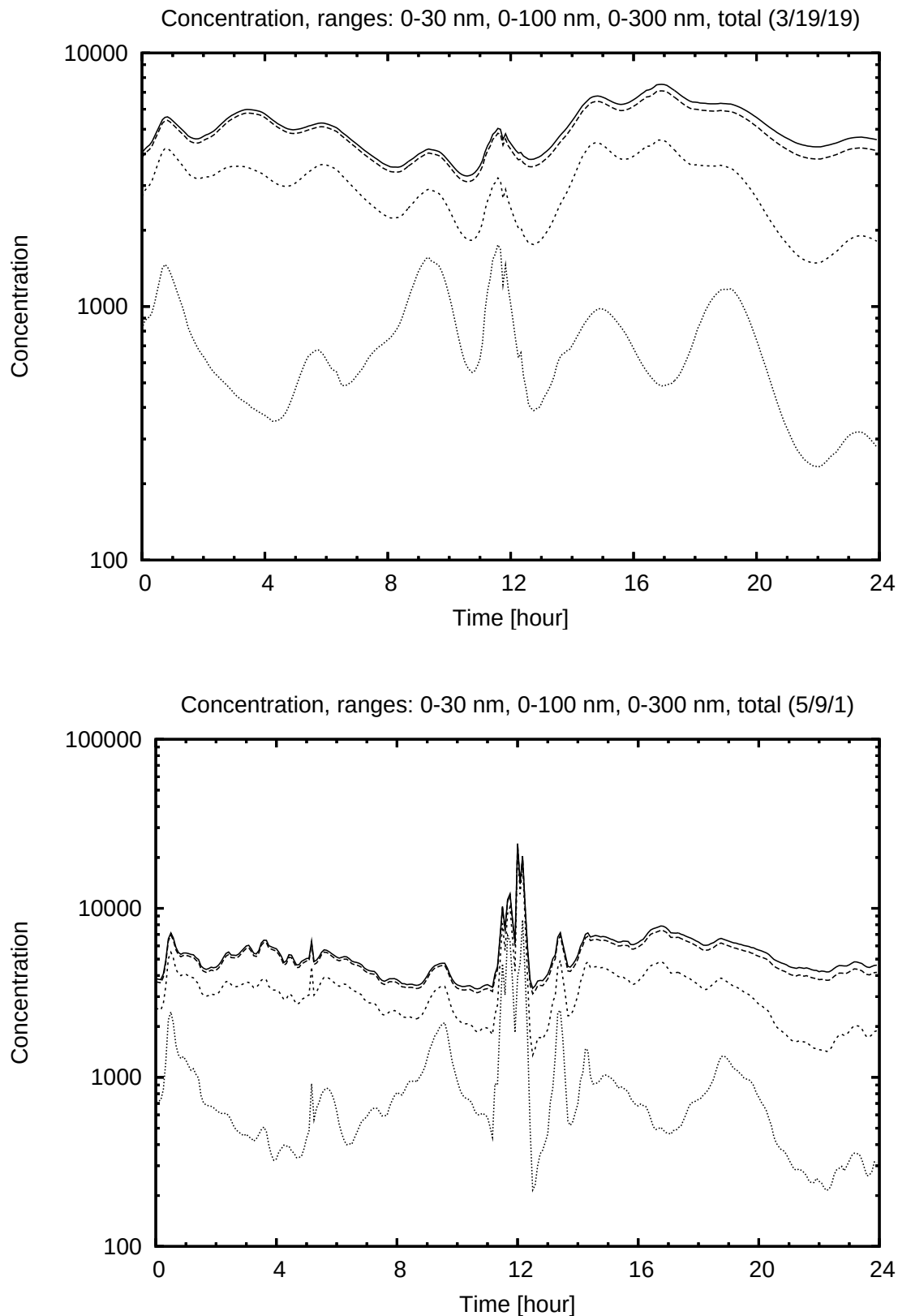
5. Softwarové řešení

Software popisovaný v této práci je postaven výhradně na programech s otevřeným zdrojovým kódem. Zpracování probíhá na operačním systému GNU/Linux, ale je přenositelné i na jiné systémy. Software byl původně vyvinut pro zpracování celých měřicích kampaní *ex post*. Výpočet bylo nutno rozložit na více paralelně běžících procesů. Řídicí skript automaticky přiděluje úlohy jednotlivým procesům. Pro tento účel byly implementovány speciální semaforey. Jsou vytvářeny atomickou operací, proto spolehlivě fungují nejen na lokálním disku, ale i na síťovém disku sdíleném prostřednictvím NFS. Programy byly dodatečně mírně upraveny, aby mohly být použity na zpracování dat v téměř reálném čase.

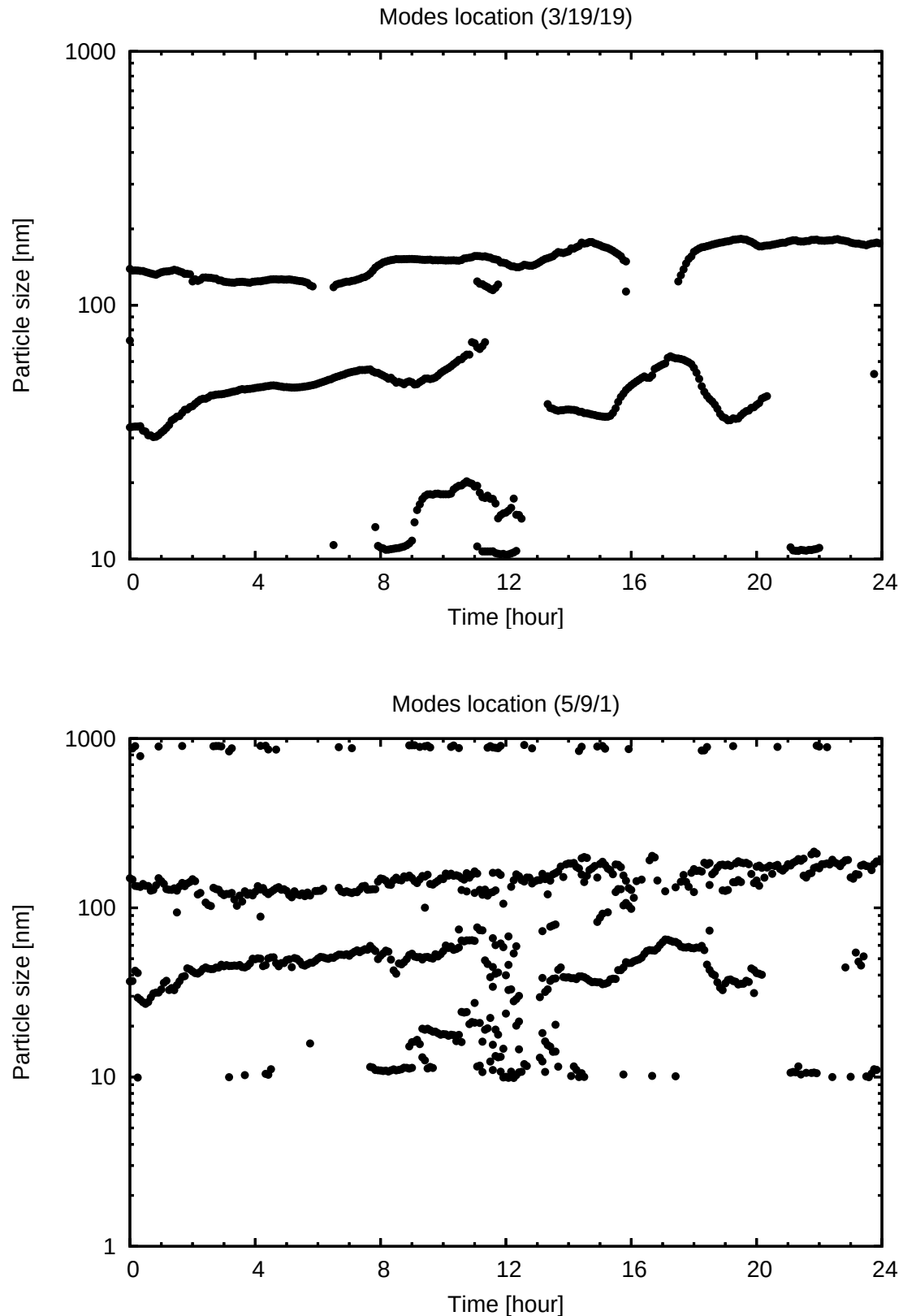
V předchozích částech bylo ukázáno, že data jsou filtrována dvěma způsoby a pro obě filtrované řady jsou stanovovány distribuční funkce velikosti



Obrázek 2: Časová závislost celkové koncentrace částic a koncentrace ve třech velikostních intervalech při výbojích na elektrodě 17. 5. 2008; srovnání dvou metod filtrace (nahore $M = 3$, $u = v = 19$, dole $M = 5$, $u = 9$, $v = 1$).



Obrázek 3: Časová závislost celkové koncentrace částic a koncentrace ve třech velikostních intervalech 8.6.2010; srovnání dvou metod filtrace při výskytu lokálního zdroje (nahore $M = 3$, $u = v = 19$, dole $M = 5$, $u = 9$, $v = 1$).



Obrázek 4: Časová závislost polohy módů 8.6.2010; srovnání dvou metod filtrace při výskytu lokálního zdroje (nahore $M = 3$, $u = v = 19$, dole $M = 5$, $u = 9$, $v = 1$).

částic. Tyto výpočty jsou na sobě nezávislé, mohou tedy být prováděny na víceprocesorovém počítači paralelně.

Zpracování je spouštěno prostřednictvím programu CRON v okamžiku, kdy mají být data z košetické observatoře již přenesena. Ke spuštění paralelních procesů a hlídání maximálně přípustného počtu paralelně spuštěných procesů je využit program GNU/MAKE. Závislosti mezi soubory nejsou jednoduché a počet souborů je velký. Navíc neustále vznikají nové soubory. Vyhodnocování závislostí programem GNU/MAKE by bylo časově náročné. Byly proto implementovány speciální časové značky, k jejichž aktualizaci se používá program TOUCH. V souboru `Makefile` jsou uvedeny pouze pseudocíle, které se sestavují vždy. Jména souborů, které s ohledem na hodnoty časových značek mají být zpracovány, se zjišťují programem FIND.

Řídicí systém je naprogramován ve skriptovacích jazycích BASH a PERL. Výpočet je prováděn programy napsanými pro OCTAVE. Grafy jsou generovány programem GNUPLOT a ukládány na disk do zvláštních adresářů. Jednoduchý skript napsaný v PHP pak vytváří webové stránky, které jsou zveřejňovány s využitím serveru APACHE.

Zpracování se provádí na počítači se čtyřjádrovým procesorem Intel s taktem 2,4 GHz a pamětí RAM 4 GiB, jehož pořizovací cena byla přibližně 15 000 Kč včetně DPH. Zpracování hodinových dat na tomto počítači trvá přibližně 15 minut. Použitý optimalizační algoritmus je stochastický, proto přesný čas výpočtu závisí na charakteru dat, ale i na zátěži počítače jinými paralelně běžícími procesy. Je tedy zřejmé, že po výpadku síťového spojení lze dosáhnout zpracování v téměř reálném čase během krátké doby.

6. Závěr

Výsledky ukazují výhody použití algoritmů matematické gnostiky na automatizované zpracování naměřených dat téměř v reálném čase. Na příkladech je demonstrováno, jak lze výsledky využít k detekci neobvyklých jevů. Je tedy možno mimo jiné rychle zjistit poruchu přístroje a včasnou opravou tak zkrátit dobu výpadku měření. To vede ke zkvalitnění časové řady měření. Systém pracuje v současné době v téměř reálném čase, což je dáno především tím, že přenos dat ze vzdálené stanice po veřejné datové síti ve skutečně reálném čase je nepraktický i z hlediska příliš velké zátěže počítače řídicího experiment. Pokud by počítač pro zpracování dat byl umístěn přímo v blízkosti měřicích přístrojů, bylo by zpracování dat v reálném čase možné.

Literatura

- [1] EUSAAR (European Supersites for Atmospheric Aerosol Research), [on-line], [cit 2011-11-30]. <http://www.eusaar.net/>
- [2] Scanning Mobility Particle Sizer Spectrometers, Preliminary Information. P/N 2980347, TSI Incorporated, 2005.
- [3] Introduction to the nasa/ames 1001 GENERAL FORMAT [on-line], [cit 2011-11-29], available only at <http://www.nilu.no/projects/ccc/eusaar/datapages/>.
- [4] Muller H.: Nonparametric regression analysis of longitudinal data. In *Lecture notes in statistics*, vol 46, Springer, New York, 1980.
- [5] Marron J. S., Chaudhuri P.: When is a feature really there? The SiZer approach. In *Automatic Target Recognition VII, Proceedings of SPIE*, vol 3371, editor Sadjadi F. A., 306–312, 1998.
- [6] Green P. J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82: 711–732, 1995.
- [7] Kovanic P., Humber M. B.: *The Economics of Information — Mathematical Gnostics for Data Analysis*, book 719 pp. [on-line]. December 2, 2009. [cit 2011-11-29], available only at <http://www.math-gnostics.com/index.php?a=books>.
- [8] Ždímal V., Brabec M., Wagner Z.: Comparison of Two Approaches to Modeling Atmospheric Aerosol Particle Size Distributions. *Aerosol and Air Quality Research*, 8(4): 392–410, 2008. http://aaqr.org/VOL8_No4_December2008/3_AAQR-08-11-0A-0051_392-410.pdf
- [9] Wagner Z., Schwarz J., Ždímal V., Eleftheriadis K., Lazaridis M., Smolík J.: Analysis of Time Evolution of Particle Size Distribution. *18th International Conference Nucleation and Atmospheric Aerosols*, Book of Abstracts, pp. 116–119, Prague, Czech Republic, 10–14 August 2009.
- [10] Heiberger R. M., Becker R. A., *J. Comp. & Graph. Stat.*, 1: 181–196, 1992.
- [11] Kovanic P., *Kybernetika*, 8: 367–383, 1972.
- [12] Kovanic P., *Kybernetika*, 10: 303–316, 1974.
- [13] Lazaridis M., Eleftheriadis K., Smolik J., Colbeck I., Kallos G., Drossinos Y., Zdimal V., Vecera Z., Mihalopoulos N., Mikuska P., Bryant C., Housiadas C., Spyridaki A., Astitha M., Havranek V., *Atmospheric Environment*, 40: 6214–6228, 2006.

APPLICATION OF SOCIAL NETWORK ANALYSIS IN BUSINESS AND SCIENCE

Eva Eckenhofer

Address: Greifensteinerstraße 113, AT-3423 St. Andrä-Wördern, Austria

E-mail: eva_eckenhofer@hotmail.com

Abstract: Social networks have attracted enormous interest in the scientific community in recent years. The characteristics, components and impacts of social networks have been studied through different kinds of aspects, such as sociological, geographical, ethnological, political and economical. Social Network Analysis (SNA) provides the theoretical background and methodology for doing so, as it constitutes a socio-anthropological method used to measure and visualize the social structure of a group as a whole and the social embeddedness of its individuals (Jansen, 2006; Wasserman & Faust, 1994). The aim of this chapter is to introduce the theoretical and analytical foundation of this method as well as providing an outline of possibilities for application in business and science. Hereby survey examples from ego-centric, socio-centric and two-mode perspective will be presented and analyzed for giving an overview of the complexity of possibilities in terms of scope of application and cognition.

Keywords: Social Network Analysis, VennMaker, Netdraw, Gephi, Visione.

1. Introduction

To form social networks is a human need and ability; from early childhood on we are members of networks, our family, school classes, sports clubs and many more. This seems to be far away from the network definition of social network analysis, where networks are described as a defined sum of nodes or elements and the sum of the edges/ties between them (Jansen, 2006). Indeed nodes are the individuals, such as persons, corporate stakeholder, companies, ministerial accounts or countries. Other authors define social networks as a defined set of persons, and the linkages between them (Tichy et al., 1979). Furthermore networks are a set of relations, which differ in aim and duration.

Various ways of distinguishing social networks exist. Basically a social network can be formal, e.g. a sports club, or informal, e.g. a group of friends. The difference between them is unproblematic; formal networks are networks whose members are listed in a certain way and the question about being a member of the network is evident, because the network borders are clearly defined. Of a sports club a list of members exists and it is easy to differentiate

who is a part of the network and who is not. For the informal network of friends it is not that clear to tell, who belongs to the group and who does not. These informal networks are classified into “communication and influence” and “exchange and negotiation” networks irrespective of their actual topic (Jansen, 2006).

A company’s network can be further distinguished into an *intra-organizational and inter-organizational network*.

Intra-organizational networks are the relations between employees, while inter-organizational networks form relations to shareholders, suppliers, costumers, competitors and any other possible stakeholder as for instance regulatory authorities. Inter-organizational networks can be further distinguished into local/global, simple/complex, obligatory/promotional/, open/closed or symmetric/asymmetric networks among others (Sydow, 2006). Moreover social networks can be divided into open and closed networks. Within open networks, which are heterogeneous, not all members know each other. Closed networks illustrate an exclusive and homogeneous network, where all members know each other (Antcliff et al., 2007).

Clusters are another type of networks, which occur, or are aimed to be built, often in nowadays economics. On one hand there are cluster areas in networks, which are more densely connected to each other (Rosen, 2000). On the other hand there are local clusters defined as geographically concentrated firms of different sizes, horizontally and/or vertically linked and operating in the same line of business (OECD, 2001).

Different positions within a network can be identified and visualized as in Figure 1. Global players are people with a lot of influence, while insiders do not have a lot of influence, but good connections to global players. Information brokers have a lot of relations, while local players only have a lot of relations in a certain branch or area (Friedschröder, 2005).

1.1. Relational Aspects

The smallest part within a network is the so called dyad, the relation between two nodes or individuals as it is shown in Figure 2. A dyad can be one-sided or reciprocal and represents a rather weak and instable relation, which can be enforced by adding more people to this relation. Relations between three nodes or individuals constitute a so called triad. Bigger sections in networks after dyad and triad are groups and cliques (Jansen, 2006).

A network differs in the structure of the relations it consists of, but furthermore the type of relation and the characteristics of this relation can be distinguished. A famous differentiation of networks is the weak and strong tie

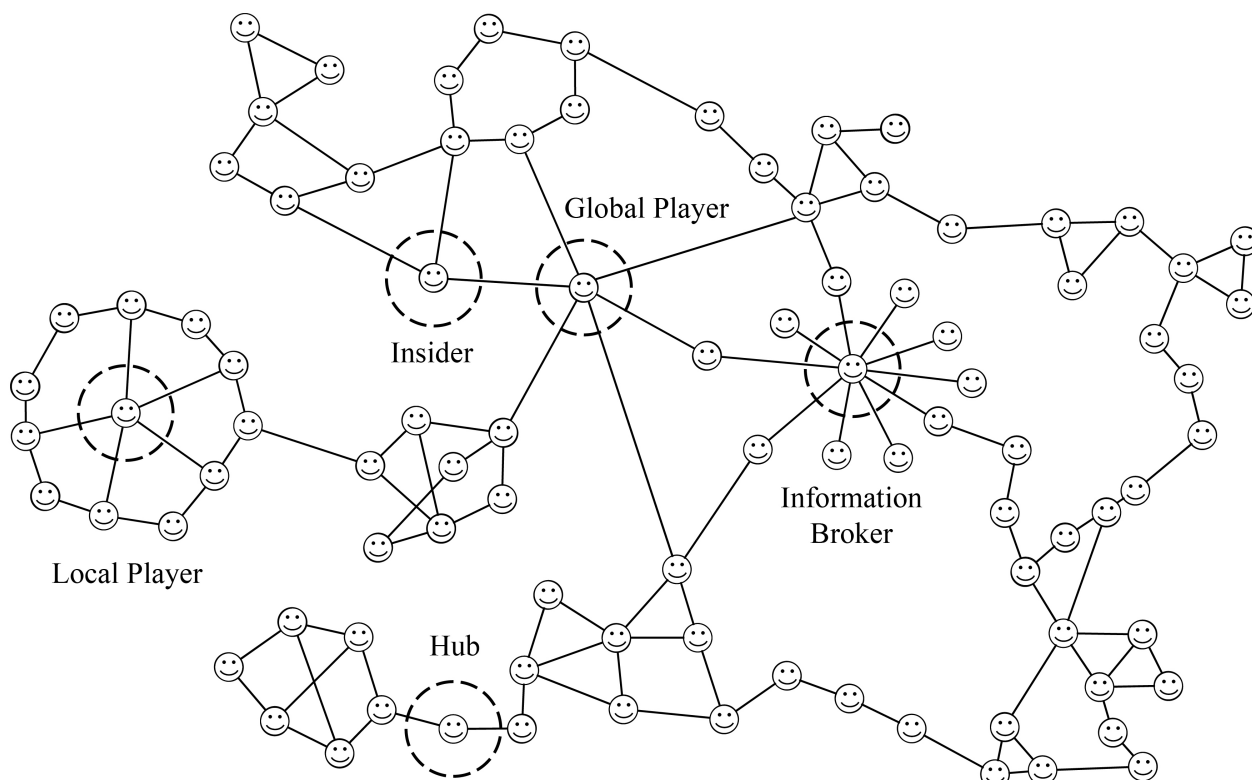


Figure 1: Positions in a Network
Source: Friedschröder, 2005, p. 70, mod.

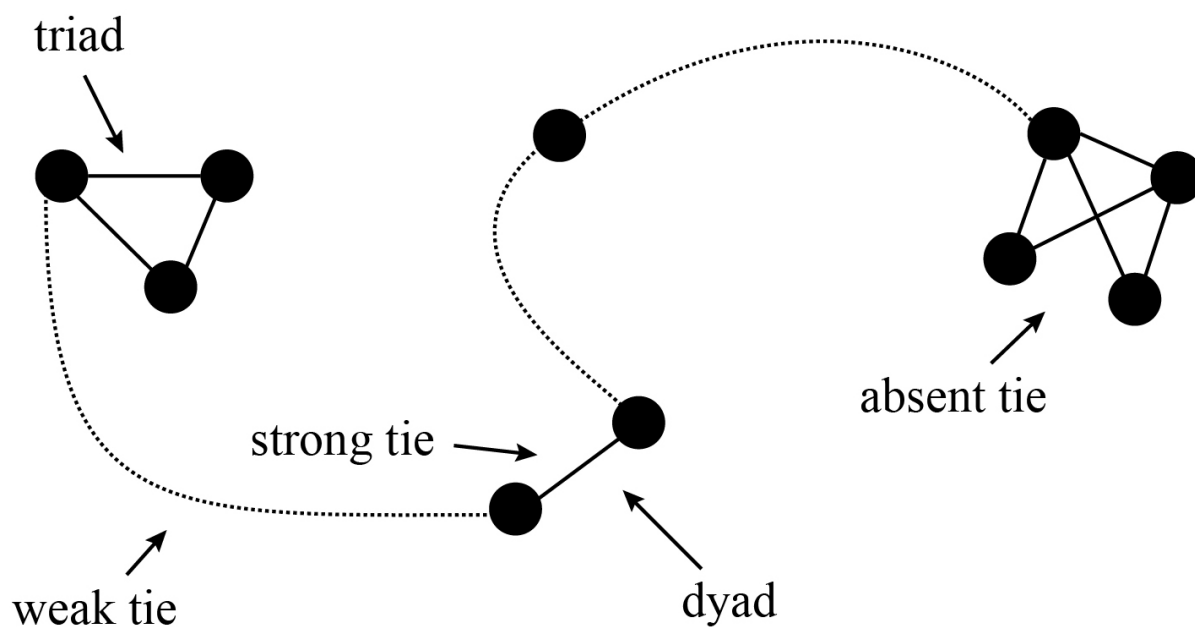


Figure 2: Ties and Nodes in a Network

differentiation by Granovetter. A graphical differentiation between weak and strong ties is shown in Figure 2. Granovetter defines a strong tie as a relation with close and intensive intercommunications, e.g. friends and families. Weak ties are defined as loose relations as to acquaintances, which have the advantage of adding new information to the network, while the advantage of strong ties is the generation of trust and solidarity (Granovetter, 1983).

The structure of social networks has been researched in various studies. The first and probably most famous study was done by Stanley Milgram in 1967. In an experiment, he proved that the average distance between people who do not know each other directly is 5.2 steps (Travers & Milgram, 1969). This is possible due to the structure of social networks, which are highly clustered, with a high density. Cohesion and density in networks provokes redundancies, a condition which leads on one hand to trust and cooperation but on the other hand to a lack of new information (Gargiulo & Benassi, 2000).

‘Small-world-networks’ are neither ordered, nor randomized. These networks, such as social networks, the internet, the human brain, street networks and many more, show ordered as well as random links, leading to a high resistance against (random) attacks. However, these networks are because of their decentralized structure easy to destroy with pointed attacks. This is possible because of their network hubs and connectors, which are part of this ‘aristocratic network’ because of preferential attachment. This structure, neither random, nor ordered, happens naturally and seems to be a natural source of security and stability (Buchanan, 2002). This structure is logical, when thinking about our social networks. Most contacts are sorted around us, neighbourhood, work, school, sports clubs, etc. But some contacts do not fit into this scheme and are acquaintances from far away. How many contacts someone has in common with his contacts can be measured and expressed by the degree of clustering (Buchanan, 2002).

An impact relating from the structure of a network is the decrease or increase of corruption. Hierarchical forms of networks lead to corruption and decrease in trust and economic development, while horizontal and more egalitarian forms of networks increase trust and economic development and moreover decrease corruption and lead to more effective governments (Halpern, 2005).

Another important aspect is complexity. More complex networks tend to fluctuate less and are more stable than simple networks (Buchanan, 2002), which seems to be highly important for firms and their intra-organizational network.

1.2. Influential Aspects

In networks there exist basically two factors of influence: Trust and Power. While trust always assumes positive consequences, power assumes negative ones. This means that in a network composed of power the individual acts because of being afraid of negative consequences, such as sanctions. Money, knowledge and democracy can be classified as a kind of power (Sydow, 2006). Trust is defined as the expectations of a partner's reliability with regard to his obligations, predictability of behaviour, and fairness in actions and negotiations and is further more a product of the successful integration of norms and values within a network (Fukuyama, 1995; Beugelsdijk & Van Schaik, 2005). Furthermore, network closure and the presence of cohesive ties promotes the development of trust (Gargiulo & Benassi, 2000).

Power appears in two ways, as legitimate power due to hierarchy, or a power due to structural holes (Burt, 1995). Control benefits from a brokerage relationship between other players. Structural holes moreover generate information benefits (Burt, 1995) and prevent amplified reciprocity. This is one kind of (negative) network consequences and leads to sanctions. Amplified reciprocity is the pressure on a person to reciprocate past favours in order not to risk gaining a tarnished reputation that may restrict the ability to make new contacts (Antcliff et al., 2007).

Reciprocity in general is a pervasive and an economically significant value in networks, no matter, if these are long-term relationships, sporadic or anonymous relations. Reciprocity is the propensity to reward generosity and punish opportunism (Sethi & Somanathan, 2003). Reciprocity is not a kind of altruism but moreover an expectation of future benefits from their action (Fehr & Gächter, 2000). From this underlying principle derives the networking principle 'Do ut des' [lat.: I give so that you may give].

Another factor in social networks is shaming as a kind of sanction. This mechanism makes social relations work, as contacts suffer the sanction of shame from their close contacts, if norms and values were offended. This goes in line with behaviour setting, a mechanism which develops common rules, cultural understandings within a social ecosystem. The members of a network, for instance a neighbourhood, school or firm, ensure these common rules and values by informing, enforcing and ejecting. This can happen in positive as well as negative means (Halpern, 2005).

Another important factor of influence within the network is the members' characteristics. Social competence is an important ability needed to cooperate successfully within networks. There are five main competences, which have been discovered to exert influence. These are social perception, the ability to

correctly gauge current moods or emotions of network partners, impression management of the own appearance and image, persuasion and influence. The ability to adjust to a wide range of social situations and to feel comfortable with individuals from diverse backgrounds (Social Adaptability) counts as well as emotional intelligence, the ability to regulate one's own emotions and have influence on the emotions of others (Beugelsdijk & Van Schaik, 2005). With this the need for proximity in social networks goes in line. Proximity is the quality of relationships in psychological, cultural, social and physical dimensions that influences the quality and quantity of communication (Becerra & Huemer, 2002), (Lechner & Dowling, 2003), (Gössling, 2007).

A factor of influence deriving from all the above mentioned factors is social capital, which is defined as the outcomes for individuals from networks with shared norms, values, and understandings that facilitate co-operation within and among groups (OECD, 2001). It is a resource embedded in social structure, which can be accessed as well as mobilized in purposive actions (Lin, 1999). As social capital is rooted in social networks and social relations it must be measured in relation to these roots and its three ingredients, the structural (embeddedness), the opportunity (accessibility) and action-oriented (use) aspects (Lin, 1999).

2. Theory of Social Network Analysis

Social network analysis (SNA) is a socio-anthropological method used to measure and visualize the social structure of a group as a whole and the social embeddedness of its individuals (Jansen, 2006; Wasserman & Faust, 1994). The evolvement of this methodology can be attributed to Jacob Moreno, who first developed the sociogram and claimed that *“before the advent of sociometry no one knew what the interpersonal structure of a group ‘precisely’ looked like”* (Moreno, 1953, p. lvi).

Social network analysis distinguishes between the analysis of an egocentric network and a socio-centric (total, whole) network. Ego-centric networks are collected from the point of view of an individual (ego), who is asked to provide his contacts according to the research question by a name generator. The names generated are called alteri and further more the relations between Ego's alteri are asked. Whole network analysis deals with all relations (according to the research question) between a defined set of actors. Social network analysis uses special software for the calculation of network measures and the graphical illustration, e.g. Ucinet, Pajek, Visone, Gephi, Netdraw, or Vennmaker. (Jansen, 2006; Schnegg & Lang, 2002).

2.1. Centrality and Prestige Measures

Social Network Analysis incorporates three main schemes to describe social network data mathematically: graph theoretic, sociometric and algebraic. In the graph theoretic scheme a relation is seen as a graph between nodes joined by lines, in the sociometric the data is represented in a two-way matrix and the algebraic is used for studying multiple relations (Wasserman & Faust, 1994).

The most important measures for characterizing social networks can be calculated for the whole network as an average of all actors. Those measures are density, describing the ratio between existing relations and all possible relations, cohesion, defined as the number of bidirectional choices in relation to the number of dyads in a network, and network multiplexity, which calculates the share of multiplex relations in all possible relations. Another important measure within networks is the measurement of the degree of homophily, which describes whether actors with similar attributes are more connected between each other than to actors with different attributes.

Other important measures for networks are centrality and prestige, concepts based on the idea that the actor, who has plenty of relations within the network, is, therefore, more central and visible. There exist three types of centrality measures of actors: degree-based, closeness-based and betweenness-based.

Degree-based centrality is measured by the outdegree of an actor, which computes all outgoing relations to other actors, in the case of an asymmetric and directional network.

$$C_{D(n_i)} = o \ d_i = \sum_j x_{ij}, \text{ for } i \neq j.$$

Closeness-based centrality measures not only the direct but moreover the indirect relations to other actors (path distances). These relations are weaker than direct relations though important as they contain a lot of new information and are easier to handle, as they do not afford a lot of time (Granovetter, 1983).

$$C_{C(n_i)} = \left(\sum_{j=1}^n d(n_i, n_j) \right)^{-1}, \text{ for } i \neq j.$$

Betweenness-based centrality follows a different logic than degree-based and closeness-based centrality as it starts from a dyad and computes the shortest path distance from one to another, called geodesic. The idea behind it is the probability that a communication from j to k will run over i . The

ratio between the number of geodesics (g) between j and k going through i to the total number of shortest paths between j and k is computed in order to get the betweenness-based centrality:

$$C_b(n_i) = \sum_{j=1}^n \sum_{k < j}^n b_{jk}(n_i), \text{ for } i \neq j \neq k;$$

$$\text{where } b_{jk}(n_i) = \frac{1}{g_{jk}} \cdot g_{jk}(n_i).$$

The centrality measures of a network show the capacity of solving problems within a group. Moreover the speed and efficiency of spreading information and solving tasks is shown by centrality. Prestige concepts cover the level of control of actors over resources and how much authority as well as attention they have in the network. (Jansen, 2006; Wasserman & Faust, 1994).

2.2. Network Modes

In social network analysis different types of social networks can be studied, according to the number of sets of actors and the properties of the ties among them. A one-mode network is a single set of actors (people, subgroups, organizations, collectives, etc.) and the relations between them. Relations can be individual evaluations such as friendship, liking, respect, or transactions, interactions, movement, formal roles or kinship. Moreover the attributes of the actors bring additional information to the relation, can be analysed in social network analysis. A two-mode network allows calculations for two sets of actors, which can be of different type, and at least one relation between them. Hereby relations between a set of actors (mode one) can be calculated through their joint affiliation with events (mode two). The nature of events can be manifold depending on the type of actors involved. Social functions can be membership in clubs, subgroups, committees or clusters. (Jansen, 2006; Wasserman & Faust, 1994).

2.3. Groups, Cliques, Roles and Positions

The target of the analysis of subgroups and cliques is to see which parts of the network are more densely connected between each other than to the rest of the network. Subgroups are classified as components, bi-components and cliques. A component is a maximal connected sub-graph, a bi-component is a cohesive group, which does not include any cutpoints or bridges. A cutpoint is a node, which if deleted splits the network in new components, a bridge

is the critical tie to the connectedness of the graph (Wasserman & Faust, 1994). A maximal cohesive subgroup of three or more actors is called clique (Hanneman & Riddle, 2005).

The clustering coefficient of an actor is a measure for calculating the openness of an actor's neighbourhood and describes a measure for the stability of the network. It describes how many contacts he has in common with his direct contacts (Buchanan, 2002). The measure was first discovered in the Small World Surveys by Watts who found out that all small world networks have a similar clustering coefficient (Watts, 1999).

Social positions and social roles are theories for describing network structure. Procedures for analyzing actors' structural similarities and patterns of relations in multi-relational networks are for instance structural equivalence, CONCOR (CONvergece of iterated CORrelations), Blockmodels and QAP (Quadratic Assignment Procedure). Those procedures split the actors according to their structural similarity or dissimilarity in groups and allow the comparison of matrices (Jansen, 2006; Wasserman & Faust, 1994).

The most important measure for calculating social roles within social networks has been proposed by Gould & Fernandez (1989). The brokerage procedure calculates measures of five kinds of brokerage: Coordinator, Consultant, Gatekeeper, Representative and Liaison. Brokerage occurs when, in a triad of the nodes A, B and C, A has a tie to B, and B has a tie to C, but A has no tie to C. That is, A needs B to reach C, and B is therefore a broker. The brokerage roles derive from membership in different groups, as displayed in Table 1:

Table 1: Overview of the Brokerage-Roles

Source: mod. Fernandez & Gould (1994), Hanneman & Riddle (2005)

Coordinator	$A \rightarrow A \rightarrow A$
Gatekeeper	$B \rightarrow A \rightarrow A$
Representative	$A \rightarrow A \rightarrow B$
Consultant	$B \rightarrow A \rightarrow B$
Liaison	$B \rightarrow A \rightarrow C$

Table 1 illustrates the composition of the five brokerage-roles, whereas the letters symbolize the membership in different groups and the arrows the direction of the ties. Another measure used for analysing brokerage within a network is the honest broker index, which measures the number of times an

actor is an honest broker, a node which is trusted by two third parties that don't trust each other (Hanneman & Riddle, 2005).

2.4. Graph Theoretical Measures of Structure

Simon (1994) claimed that hierarchy exists in any system, even in informal organizations, as it allows the system to operate more efficiently and survive disturbances. He argued that as informal organizations evolve naturally they follow the trend that communication flows tend to centralize, even though they might have a flat hierarchy disregarding the official communication patterns. Deriving from that idea Krackhardt (1994) proposed four measures for evaluating a social network's hierarchy. Those measures are the connectedness, the hierarchy, the efficiency and the least upper bound. The idea of those measures is to compare a given social network to an outtree, an archetype of a perfect hierarchical system.

All of Krackhardt's measures are based on the number of outtree violations calculating a continuously varying value from 0 to 1. The *degree of connectedness* is defined as the number of violations of the connectedness condition, whereas a violation is defined as two points unable to reach each other. The *degree of hierarchy* is defined as the number of violations against the reachability digraph, and measures the extent to which paths are not reciprocated. The idea is that in an outtree (such as the organizational chart), a subordinate cannot be the boss of a boss. The *degree of graph efficiency* measures the extent of existing redundant ties, which are not bridges and where a deleting of the ties does not lead to a splitting of a component. The idea behind it is that links are not without costs in a social system, and take time and resources. Therefore a social network is more efficient if it only consists of necessary relations. Therefore graph efficiency reflects the cost of a dense network. The *degree of least upper bound (LUB)* measures whether every pair has access to a common third person in the organization to whom they both can "appeal", which is a measure for conflict resistance in a network (see Figure 2). In an outtree the calculation of all above presented measures gives a value of one (Krackhardt, 1994).

3. Survey from Ego-centric perspective

The collection and evaluation of inter-organizational networks and relations has tradition in business research already since the 90ies of the last century (Saxenian, 1991; Provan & Milward, 1995, 2001; Borgatti & Foster, 2003). Especially the collection of ego-centric networks is a popular approach towards

research of organizational networks; although lately the analysis of whole networks became more fashionable, especially as it was strongly promoted by some scientists (Provan & Sydow, 2008). The choice of appropriate level of analysis and the definition of the relevant relation types was a problem ever since. In order to overcome this difficulty the stakeholder concept was taken as a background for the definition of the appropriate level of analysis in this specific survey. A stakeholder is defined as “*any group or individual who can affect or is affected by the achievement of the organization’s objectives*” (Freeman, 1984, p. 46) and stands in contrast to the Stockholder/Shareholder approach (Smith, 2003).

The six main stakeholder groups, customers, partners and suppliers, competitors, public administration/government, public/media/NPO’s and shareholder have been selected for collecting the informal (not institutionalized), inter-organizational network to them using the software VennMaker (Kronenwett & Schönhuth, 2011).

In a participative expert interview with the management of a firm, by name generator (Wolf, 2006; Hennig, 2008) contacts in each stakeholder-group have been collected and in a next step linked to the participating firm, which is in the centre of the network. Four relational types were available: formal, informal, trustful, and critical, whereby informal and formal, describes the communication type and not whether the relationship is institutionalized. The generated stakeholder-networks allowed an evaluation and assessment on their effectiveness and efficiency using social network analysis. This research project with the number IGA/49/FaME/11/D supported by the Internal Grant Agency of Tomas Bata University, covered in total eight stakeholder networks. In the following one of those networks will be presented and analyzed.

Company 1, a supplier of software solutions with 48 employees, which is for 25 years in the market and located in southern part of Austria, has in total 66 actors in its network. 25 are in the closest circle, 23 are the medium one and to 18 actors Company 1 has a weak relationship. The most actors can be found in the sector partners and supplier (21), followed by 18 in the sector customer. As the least number of actors can be found in the sector public/media/NPO, it can be concluded that Company 1 has a weak focus on public relations (PR), see Figure 3.

The overall density is low; with ego only 5% of all possible relations are realized. Without ego, only 2% are realized between the alteri. Here it has to be outlined that an ego-centred perspective of network analysis only collects and analyses the relations as perceived and known by ego (Diaz-Bone, 1997). Therefore it can happen that more relations between the alteri exist in reality

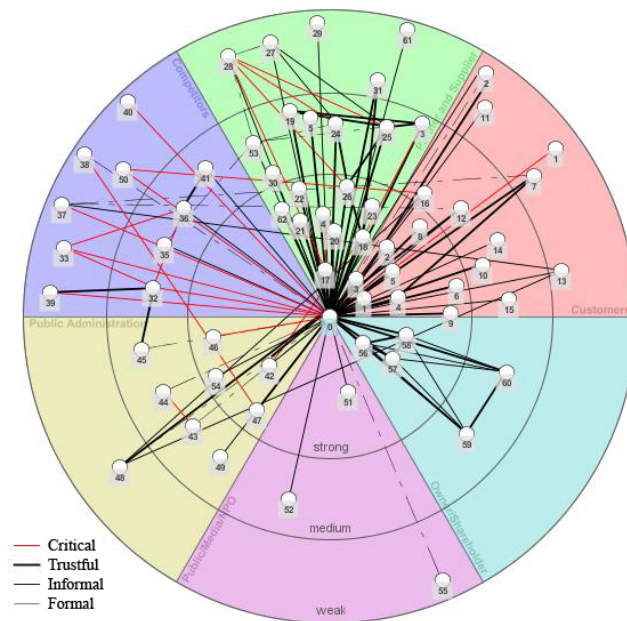


Figure 3: Stakeholder Network of Company 1
Created in VennMaker.

than ego is informed about, though usually it can be expected that till the second degree ego is aware of alteri knowing each other or not (Granovetter, 1973). Ego's centrality in the network is with the betweenness value of 0.967 compared to the other companies in the highest. This constellation of low density and high centrality provides advantages to the focal firm as it is due to this network structure a commander in his network and can by that resist stakeholder pressures while having influence on the information flow (Rowley, 1997).

Concerning the relationship types 40% of the relations of Company 1 are trustful. 29% are informal, 12% formal and 19% critical. Most of the critical relations are to competitors of the firm, though also to one partner and one customer. The critical relations between the partners and suppliers lead to an advantageous position of ego as a *'tertius gaudens'*, the laughing third (Burt, 1995). The balanced relational mix concerning relational types as well as proximity provides Company 1 advantages of neither being under-embedded nor over-embedded in their network (Uzzi, 1997).

Nevertheless shows brokerage analysis that several critical brokerage positions can be found, which might in worst case harm the company's success. Those are several very strong coordinators in the stakeholder group of partners and suppliers, as well as consultants and gatekeeper in the group of competitors. Twice a liason has been found in the group of customers. Those actors which have been found to be a broker, have to be observed and their

power kept in mind by the firm. Moreover harmful can be also those broker which are unknown by the firm, but might be identified by the theory of transitivity as well as the small world structure of a network.

Analysis of the graph theoretical structure of the ego-centric network showed that the connectedness is 1, no alteri is without relation to ego, hierarchy is 0, which is given due to the network type, Least Upper Bound 1 and network efficiency 97.86%, which tells us that almost no relations in this network are without function to the network and that the network is cost-efficient for the firm.

4. Survey from Socio-centric perspective

In this chapter an example for socio-centric social network analysis will be shown. This is a data set that focuses on the intra-organizational relations between employees of the same company, whose stakeholder network has been presented and analyzed in Chapter 3.

The data about the social networks has been collected via online questionnaire, which were addressed to all members of the defined group, to all employees. In order to calculate valid whole networks, where a participation of a minimum of 70 percent of a defined group was needed (Schnegg & Lang, 2002). The questionnaire for employees offered seven types of informal relationships with other employees of the company: Talking regularly about business topics (Relation 1), working on joint projects (Relation 2), asking for professional advice (Relation 3), talking about private topics (Relation 4), meeting in free time (Relation 5), asking about private advice (Relation 6), and the probability of lending 200 Euro (Relation 7) as well as trust concerned and demographic questions.

The social networks of Company 1 have an overall high density and are well connected. An exception builds relation 5 and 6, private advice and private meetings, where the degree of connectedness drops on 0.28 and 0.45. The degree centralization of the aggregated network is with 17.2% low as well as the brokerage. In the aggregated network exist only 5% of “honest broker”, which are actors having ties to not connected alteri. Moreover is the average brokerage according to Fernandez and Gould’s brokerage positions with 5.79 low.

In the following Relation 1 (Professional Talk), Relation 3 (Professional Advice) and Relation 7 (Lending Money) will be presented and discussed on the basis of their network graphs.

Figure 4 shows the social network graph of the first relation in Company 1, the relation of communication with colleagues about professional topics. The

graph-theoretical layout of this network, as well as all following ones, was generated by spring embedding, an algorithm that uses iterative fitting to locate the points to each other according to their smallest geodesic distance (Trappmann et al., 2005). The colour of the nodes has been selected according to their department, which means that nodes with the same colour belong to the same department, which makes a graphical analysis of homophily between departments possible. Important is hereby that the management has black node colour. The node size has been chosen according to the betweenness value of an actor, which expresses information benefits.

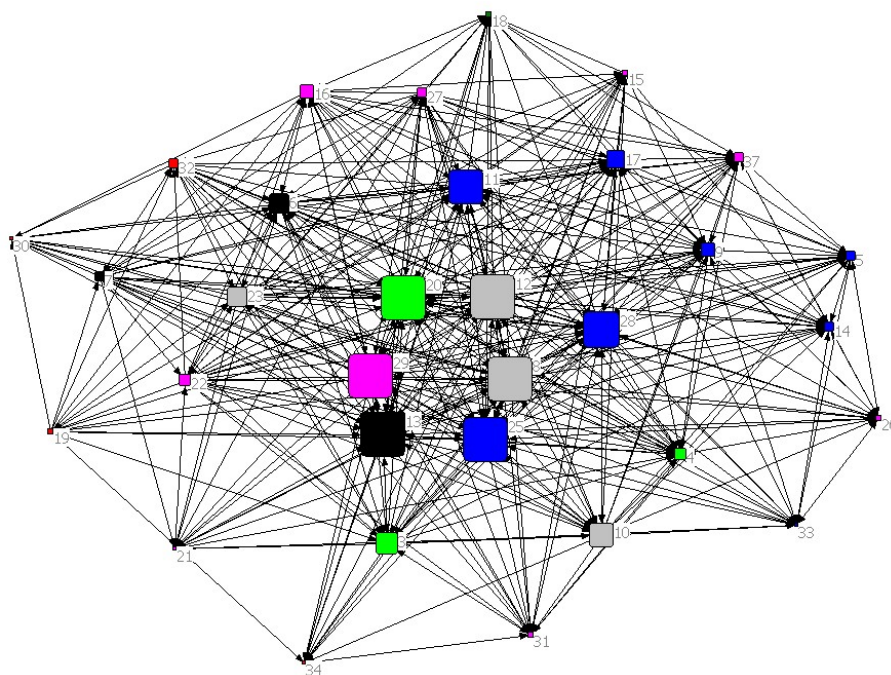


Figure 4: Relation 1 (Professional Talk) of Company 1
Created by Netdraw.

The talking network of Company 1 consists of one component and shows a high density; almost 50% of all possible relations are realized. The average path length is 1.56, which means that it takes on average 1.5 steps to reach any other contact in the network, which can be considered as a positive value as this characteristic allows a good flow of information. Moreover there are no signs of homophily in this relational network as the homophily index E-1 index is 0.59, whereas on a scale from -1 to $+1$, -1 means homophily and $+1$ heterophily. A few actors are more central in the network and also inhibit higher betweenness values than others, though in general the centrality is balanced. Also one member of the management with the number 13 is central

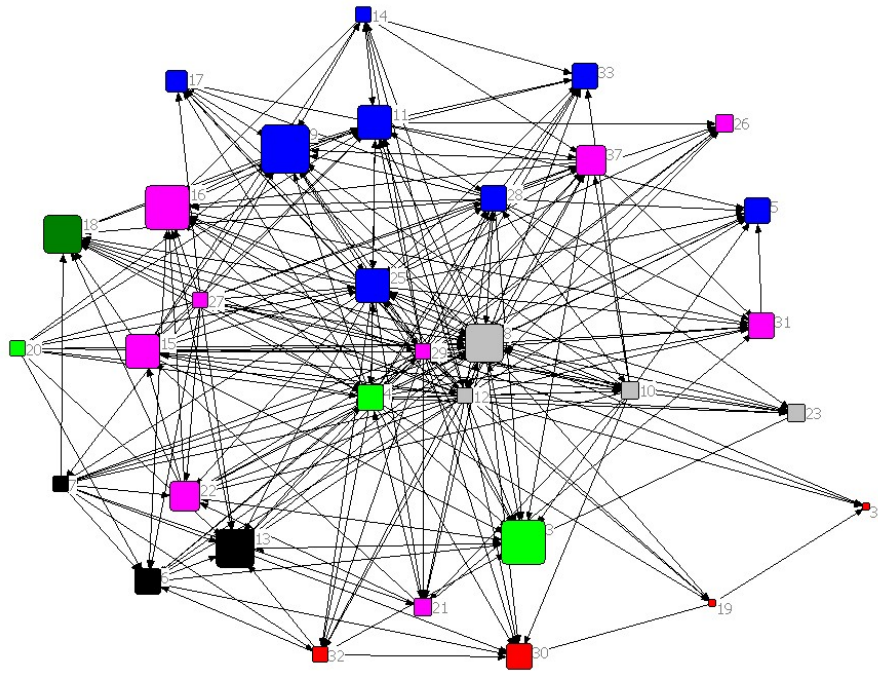


Figure 5: Relation 3 (Professional Advice) of Company 1
Created by Netdraw.

in the network, while the other two members of the management are more in the periphery and close to each other (7, 5).

Figure 5 shows Relation 3 of Company 1, the network of professional advice seeking, which consists of one component and a degree centralization of 56%. The difference between Figure 5 and Figure 4, the talking relation, is big. This network has a smaller density of 28% and also the path length is higher, 1.79. In the network in Figure 5 the node size has been set according to the indegree value of the individual actor. A high indegree value means that many of their colleagues indicated that they would ask this actor for advice in professional matters. It can be observed that in the network of advice seeking a few (4–5) actors are in the centre of the network due to their outdegree, though relatively small in terms of indegree. The management of Company 1 (5, 7, 13) builds a triad at the left periphery of the network. That the management is not central in this network is not surprising as when people move higher in an organization, due to more administrative tasks, they get less accessible and less knowledgeable about the day to day work of their subordinates (Cross et al., 2001). Leading actors in terms of indegree are actor 9, 3 and 16.

Figure 6 shows the last relation, which was generated by asking “Whom would you lend an amount of 200 €”. This questions is testing the trust within

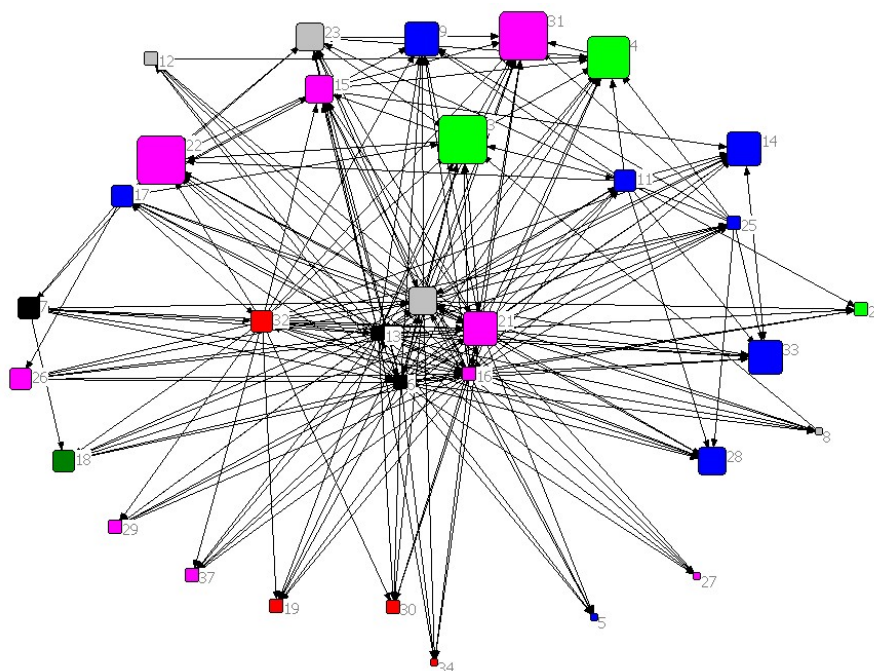


Figure 6: Relation 7 (Lending Money) of Company 1
Created by Netdraw.

a social network, as lending an amount of 200 € already requires a minimum of trust.

Six actors are in the centre of the social network and build a clique by lending money to each other. Though, the node size, which has been set by the indegree of an actor, shows that those actors are not the most trusted ones in the network. Those actors, enjoying a high level of trust and have an indegree of 11 and are located at the upper periphery of the network with the numbers 3, 22 and 31. On the right side of the figure a group of actors of the same department can be found, lending money to each other. This shows that the level of trust in this particular department is high. The density of the network is 23% with an average path length of 1.6. This goes in line with the general trust evaluation, where 57.1% indicated that in general most people can be trusted. Just 28.6% answered that in general you cannot be too careful. The rest answered with “I don’t know” or gave another answer.

Concerning multiplexity it can be stated that more than the half of the relations (64%) of the relations are multidimensional and provide the network therefore stability and continuity. The majority of the multiplex relations span over two (22%), three (23%) or four relations (10.5%), though still there can be found 5% of the relations which are connected over five relational

types, 3% that span over 6 relations and 0.5% relations that go over all seven relational types.

5. Two-mode surveys

The following example for two-mode network analysis has been taken from the sphere of clusters. Clusters are the most prominent and important example for regional networks, which have enjoyed enormous attention in recent years, as being a valve for regional competitiveness and innovation (OECD, 2007), and consist typically of small and medium sized companies in a local agglomeration (Sydow, 2006). Clusters have been described as critical masses in one place that show an unusual success in a particular field (Porter, 1998). Thirty actively operating clusters located in the Czech Republic have been taken and a database with all their member firms generated. In a next step two-mode network analysis has been applied and the member companies linked to the clusters by their membership in order to see whether double membership and the integration of local cluster agencies and cooperation with universities lead to indirect relations between clusters. Figure 7 shows the graphical result of this step.

The graph-theoretical layout of the network was generated by spring embedding, and therefore nodes are arranged according to their smallest

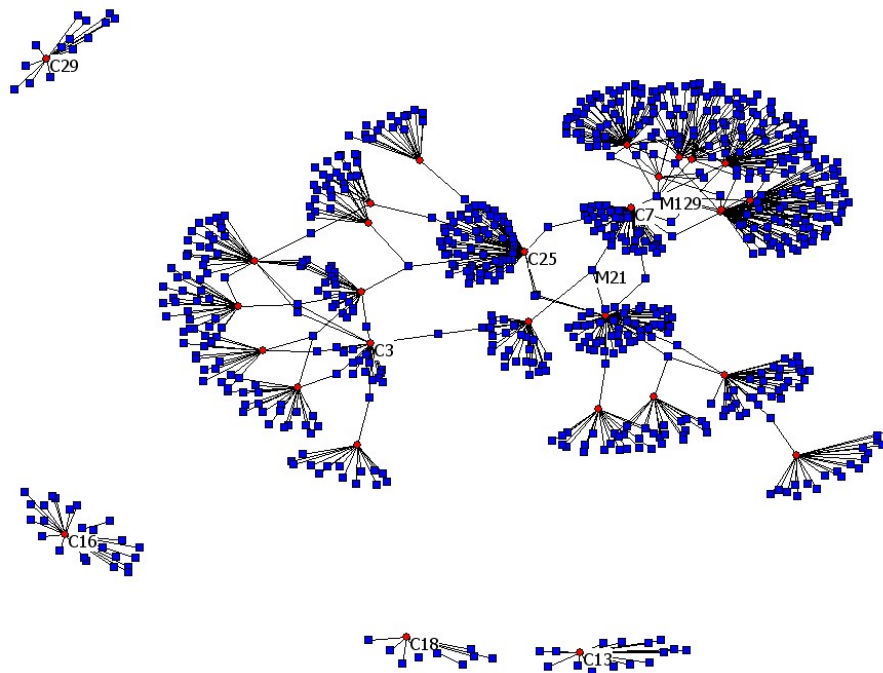


Figure 7: Two-Mode Cluster-Member Network
Created by Netdraw.

geodesic distance (Trappmann et al. 2005). Red nodes in the network represent clusters and blue nodes are members of clusters. The graph in Figure 7 shows moreover that a big group is connected to each other and builds one component with 708 nodes. Four more components exist in the network, which are neither connected to the big component nor to any other. Those components have between 10 and 25 nodes. The density of this network is 0.036 indicating that 3% of all possible relations are realized. The average distance between nodes is 6.279 and the fragmentation of the network, indicating the disconnectedness, 0.161. The longest geodesic path in the network measures twelve edges.

Analysis of the centrality of actors shows which actors are powerful in the network. According the degree, the sum of outgoing relations, C25 (Klastr obecného strojírenství) has a leading role. This can be explained by the number of members this cluster has, as it is the biggest cluster in the sample according to the number of its members (76 members).

In matters of betweenness, a ratio indicating how many times an actor lies on the direct path between two others, the actors M129 (Sdružení pro rozvoj Moravskoslezského kraje) and C7 (Moravskoslezský automobilový klastr) enrol a leading position, which gives an insight about information advantages due to strategic positions in the network. Actor M129 (Sdružení pro rozvoj Moravskoslezského kraje) is a regional institution, which cooperates with all clusters in the Moravian-Silesian Region. Actor C7 (Moravskoslezský automobilový klastr) has the largest number of members (5), which are connected to other cluster at the same time.

Closeness-based centrality measures the direct and the indirect relations to other actors (path distances). These relations are weaker than direct relations though are important as they imply a lot of new information and are easier to handle, as they do not afford a lot of time (Granovetter, 1983). These advantages result for C25 (Klastr obecného strojírenství) and M21 (Smartplast, s. r. o.) who have high closeness values.

By adding another network mode, cooperation with universities, a more complex picture can be shown, which is illustrated in Figure 8.

The nodes in Figure 8 are arranged by spring embedder and the colours of the node show the type of actor. Clusters are coloured in red, members in black and universities in green. The density of the cluster-member-university network is 0.037, with a fragmentation of 0.027 and an average distance of 5.394. This shows that the universities are the connecting item in the network creating a higher density to the cluster-member network and lead to a considerable decrease of the fragmentation of the network. Only one cluster

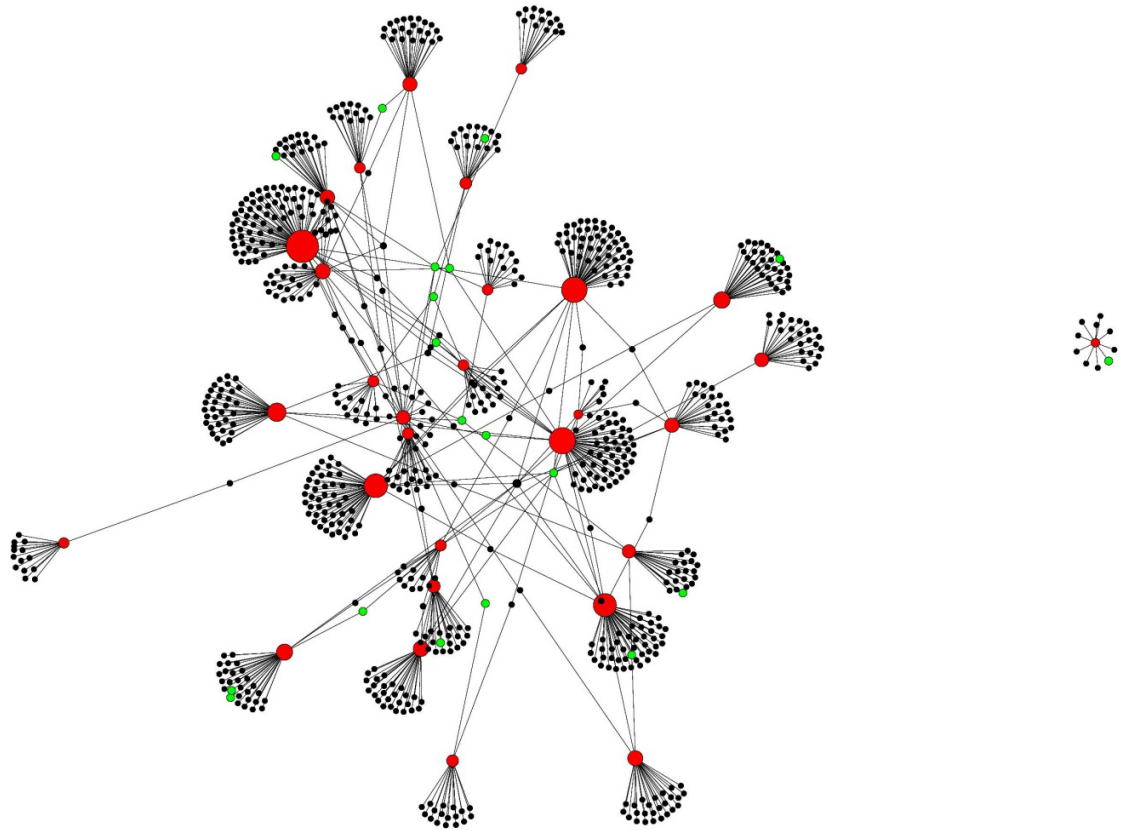


Figure 8: Network of Clusters, Members and Universities
Created by Gephi.

is left that is not linked to the large component in the middle, though has an active cooperation with a university.

A simpler example of two mode network analysis is the analysis of students' attendance sheets via two-mode analysis, which allows identifying students that have visited lectures together. Moreover the most popular lectures as well as most active students can be identified graphically. In some cases also irregularities in the attendance sheet can be visualized.

Figure 9 shows a group of 18 students, subscribed to a lecture that took place eight times. The students are illustrated by red round nodes and the lectures by blue squared nodes. The nodes have been arranged by metric MDS (Multi-Dimensional Scaling), a procedure that allocates the nodes according to their proximity in terms of Euclidean distance for instance. In our example MDS arranged students visiting always the same lectures on top of each other. An example is the group of students TT, CJ and TM, who not only went always together, to the lectures, but moreover went to most of the lectures. The most popular terms were March 4th and March 11th.

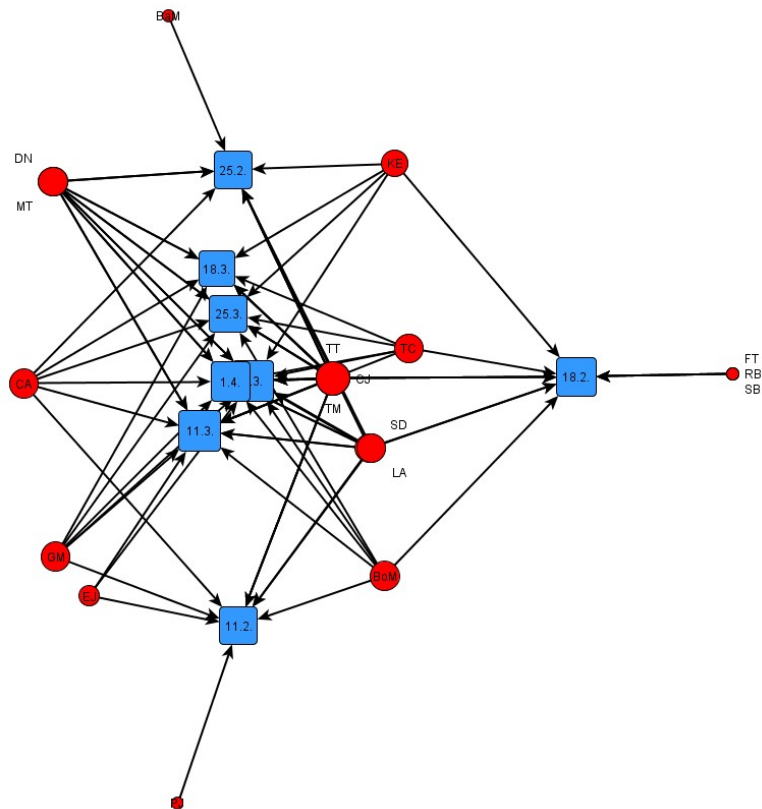


Figure 9: Bi-modal network of Students visiting lectures
Data provided by Pavel Stříž. Created by Visone.

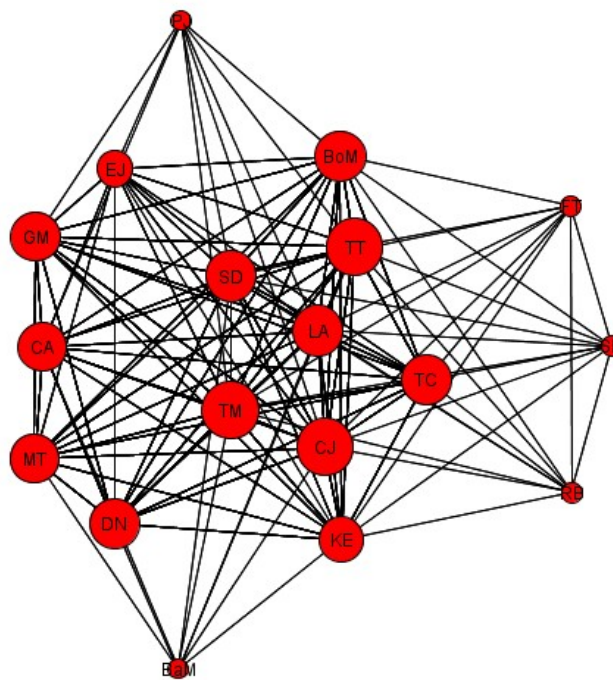


Figure 10: Affiliation network of students
Data provided by Pavel Stříž. Created by Visone.

In a next step an affiliation, can be done, to see which students are connected to each other via the lectures and a one-mode network is calculated from the rows (students) of the two-mode network and the graphical visualisation displayed in Figure 10.

The graphic shows that all students theoretically know each other, as according to the attendance sheet they have met in the lecture. The node size is chosen according to the degree of an actor and shows the students, which have not been regularly in the lecture with a smaller node size. In the centre of the network are those students, which have been in many lectures and therefore “know” most of their colleagues.

This procedure can be applied for various purposes. Joint membership or attendance in clusters such as above, but also sports clubs, alliances, conferences, joint-publication, and many more. Actors can also be affiliated because of the region the live/operate in, land/items they own or confessions they share, according to the (research) question posed.

6. Conclusion

After Introduction about the nature and characteristics of social networks, this article introduced the theoretical aspects behind social network analysis and the most common procedures, which can be applied in practical research. Chapters 3, 4 and 5, presented different research samples, where from a different point of view social network analysis was applied. In Chapter 3 an ego-centric perspective was chosen, in Chapter 4 a socio-centric one and in Chapter 5 bi-modal data has been affiliated.

The examples, taken from scientific data, collected and analysed by the author in prior research, showed how manifold the methodology of social network analysis can be applied. Not only that social network analysis allows to analyse networks from different angles, but moreover the data can be of various nature. In Chapter 3 relations between stakeholders of a firm have been analysed, in Chapter 4 relations between employees of a firm and in Chapter 5 relations between clusters and their members as well as students and lectures. This shows that the actors, can be individuals as well as groups or institutions. Not only the nature of the actors can vary, but moreover the ties that connect the actors.

In the first sample, from ego-centric perspective, we were looking at communication ties, which were either formal, informal, trustful or critical. The nature of the relations in the socio-centric example of employees of a firm, were defined by an action. The action for relation 1 was: Talking about professional topics; for relation 2: Working on a joint project; and in relation 6:

Lending an amount of 200 €. Different in Chapter 5, here the relations of the bi-modal networks were formal and institutionalized. In the first example of two-mode networks, the relations were defined based on the membership of companies in a cluster and in the second example the relations were based on the formal attendance sheet of students subscribed to a lecture.

Hereby this article presented only a few possibilities for applying social network analysis is socio-economic research. Though the opportunities are not limited to this aspect. Social network analysis can be applied in the field of geography, biology, history, ethnology, politics and many more.

With the numerous possibilities, that social network analysis offers, also a few problems go in line with it. Due to the development of theory and methodology of social network analysis in different fields, e.g. mathematics, sociology; it lacks of affiliation to a certain field and a consistent theory (Jansen, 2004; Wasserman & Faust, 1994). Another difficulty which appears when doing social network analysis, is the sensitivity of the data, which makes data collection problematic due to data privacy matters and therefore access to data limited. The data from the presented survey has been collected by different approaches. The data from Chapter 3, has been collected by personal interviews with the management of the firm. This approach allowed direct anonymization by the firm in case needed. Moreover questions that arose during the interview, could have been answered directly and lead to more clarity. The application of online questionnaires, which was used for the data presented in Chapter 4, solved the difficulty of data privacy only indirect.

Online questionnaire partly gave the participant the feeling of anonymity, though to some extent the participants had more doubts about data security. A big advantage of the use of online questionnaires was definitely that the data is available digitally afterwards, which saves a lot of time and limits sources of errors from the data input. Nevertheless it adds one source of error as during an oral interview, the interviewer can check whether the interviewee has understood the questions and is answering honestly. The data from the two-mode networks derives from desktop survey. A database of all institutionalized clusters and their members has been generated from the websites of the corresponding clusters. This approach is definitely, the easiest approach, even though being time consuming.

The opportunities from applying social network analysis deriving from data available on the World Wide Web are huge. Here social networks, where users are sharing voluntarily a lot of information, are buzz research field at the moment, where still a lot of questions are not answered yet.

Useful if not, necessary is the knowledge and expertise in the use of social network analysis software, that have been developed and improved in the

last years. Those programs, partly freeware, allow the implementation of the data as matrices or edge lists, the analysis of the data and the visualisation of the latter. The most famous programs are Pajek and Ucinet. While Pajek is freeware and handy for large networks, Ucinet is a commercial product developed by Lin Freeman, Martin Everett and Steve Borgatti. Other software applied in this article is Visone, useful for calculation and visualization of smaller networks, Gephi, developed for the visualization of large networks and VennMaker, which is developed for the data collection of ego-centred networks and includes only a few calculations. (Below the list of references a list of social network analysis software is given.)

This article aimed to provide a broad overview about social network analysis as a theory and methodology and tried to show the reader the broad range of opportunities. For more information shall be referred to Jansen (2006), Wasserman & Faust (1994) as well as the website of the International Network for Social Network Analysis (INSNA, <http://www.insna.org/>).

References

- [1] Antcliff V., Soundry R., Stuart M.: Networks and Social Capital in the UK Television Industry. The Weakness of Weak Ties. *Human Relations*, 60(2): 371–393, 2007.
- [2] Becerra M., Huemer L.: Moral Character and Relationship Effectiveness: an Empirical Investigation of Trust within Organizations. In von Weltzien Hoivik, H. *Moral Leadership in Action*. Building and Sustaining Moral Competence in European Organisations. 1st ed. Cheltenham: Edward Elgar, 2002, pp. 71–83.
- [3] Beugelsdijk S., Van Schaik T.: Differences in Social Capital between 54 Western European Regions. *Regional Studies*, 39(8): 1053–1064, 2005.
- [4] Borgatti S., Foster P.: The Network Paradigm in Organizational Research. A Review and Typology. *Journal of Management*, 29(6): 991–1013, 2003.
- [5] Buchanan M.: *Nexus*. Small Worlds and the Groundbreaking Science of Networks. 1st edition. New York: W. W. Norton, 2002. 256 p. ISBN 0393041530.
- [6] Burt R. S.: *Structural Holes*. The Social Structure of Competition. 1st paperback edition. Cambridge, MA, London: Harvard University Press, 1995. 313 p. ISBN 0674843711.
- [7] Diaz-Bone R.: *Ego-zentrierte Netzwerkanalyse und familiäre Beziehungssysteme*. 1st edition. Wiesbaden: Deutscher Universitätsverlag, 1997. 236 pages. ISBN 3-8244-4208-6.
- [8] Fehr E., Gächter S.: Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, 14: 159–181, 2000.

- [9] Freeman E.: *Strategic Management*. A Stakeholder Approach. 1st edition. Boston: Pitman, 1984. 276 pages. Pitman Series in Business and Public Policy. ISBN 0273019139.
- [10] Friedrichschröder T.: *Spielanleitung für Netzwerker*. Beruflichen Erfolg und Karriere bewusst steuern. 1st edition. Austria, Vienna: WUV, 2005. 149 pages. ISBN 3851149211.
- [11] Fukuyama F.: *Trust*. The Social Virtues and the Creation of Prosperity. 1st edition. New York, London: Free Press, 1995. 458 p. ISBN 0029109760.
- [12] Gargiulo M., Benassi M.: Trapped in Your Own Net? Network Cohesion, Structural Holes, and the Adaption of Social Capital. *Organization Science*, 11(2): 183–196, 2000.
- [13] Gössling T.: Proximity, Trust and Morality in Networks. *European Planning Studies*, 12(5): 675–689, 2007.
- [14] Gould R., Fernandez R.: Structures of Mediation. A Formal Approach to Brokerage in Transaction Networks. *Sociological Methodology*, 19: 89–126, 1989.
- [15] Granovetter M.: The Strength of Weak Ties. A Network Theory Revisited. *Sociological Theory*, 1: 201–233, 1983.
- [16] Granovetter M.: The Strength of Weak Ties. *American Journal of Sociology*, 78(6): 1360–1380, 1973.
- [17] Halpern D.: *Social Capital*. 1st edition. Cambridge: Polity Press, 2005. 388 p. ISBN 0-7456-2547-9.
- [18] Hanneman R. A., Riddle M.: *Introduction to social network methods* [online]. 1st edition. [Riverside, CA (USA)]: University of California 2005 [retrieved: 13. 08. 2011]. Available on World Wide Web: <http://faculty.ucr.edu/~hanneman/nettext/>
- [19] Hennig M.: Mit welchem Ziel werden bestehende Netzwerke generiert? In Stegbauer, C. *Netzwerkanalyse und Netzwerktheorie*. Ein neues Paradigma in den Sozialwissenschaften. 1st edition. Wiesbaden: VS Verlag für Sozialwissenschaften, 2008, pp. 295–307.
- [20] Jansen D.: *Einführung in die Netzwerkanalyse*. Grundlagen, Methoden, Forschungsbeispiele. 3rd edition. Wiesbaden: VS Verlag für Sozialwissenschaften, 2006. 312 p. ISBN 3531150545.
- [21] Krackhardt D.: Graph Theoretical Dimensions of Informal Organizations. In Carley, K. M.; Prietula, M. J. *Computational Organizational Theory*. 1st ed. Hillsdale, New Jersey: Erlbaum, 1994, pp. 89–111.
- [22] Kronenwett M., Schönhuth M.: *VennMaker 1.2*. Anwenderhandbuch [online]. 1st edition. [Trier (Germany)]: University of Trier, June 2011 [retrieved: 21. 08. 2011]. Available on World Wide Web: http://vennmaker.uni-trier.de/dl/VennMaker_1.2_Anwenderhandbuch.pdf

- [23] Lechner C., Dowling M.: Firm Networks. External Relationships as Sources for the Growth and Competitiveness of Entrepreneurial Firms. *Entrepreneurship & Regional Development*, 15(1): 1–26, 2003.
- [24] Lin N.: Building a Network Theory of Social Capital. *Connections*, 22(1): 28–51, 1999.
- [25] Moreno J.L.: *Who Shall Survive?* Foundations of Sociometry, Group Psychotherapy and Sociodrama. 2nd edition. Beacon, New York: Beacon House, 1953. 763 p.
- [26] Organisation for Economic Co-Operation and Development (OECD): *Competitive Regional Clusters*. National Policy Approaches. 1st edition. Paris: OECD Publication, 2007. 350 pages. OECD Reviews of Regional Innovation. ISBN 9789264031821.
- [27] Organisation for Economic Co-Operation and Development (OECD): *Enhancing SME Competitiveness*. The OECD Bologna Ministerial Conference. 1st ed. Paris: OECD Publication, 2001. 268 p. OECD Proceedings. ISBN 9264186492.
- [28] Porter M. E.: Clusters and the New Economics of Competition. *Harvard Business Review*, 11/12: 77–90, 1998.
- [29] Provan K. G., Milward H. B.: A Preliminary Theory of Network Effectiveness. A Comparative Study of Four Community Mental Health Systems. *Administrative Science Quarterly*, 40(1): 1–33, 1995.
- [30] Provan K., Milward H. B.: Do Networks Really Work? A Framework for Evaluating Public-Sector Organizational Networks. *Public Administration Review*, 61(4): 414–423, 2001.
- [31] Rosen E.: *The Anatomy of Buzz*. How to Create Word-of-mouth Marketing. 1st edition. New York: Doubleday Currency, 2000. 303 p. ISBN 0385496672.
- [32] Rowley T.: Moving Beyond Dyadic Ties. A Network Theory of Stakeholder Influences. *Academy of Management Review*, 22(4): 887–910, 1997.
- [33] Saxenian A.: The Origins and Dynamics of Production Networks in Silicon Valley. *Research Policy*, 20(5): 423–437, 1991.
- [34] Schnegg M., Lang H.: *Netzwerkanalyse*. Eine praxisorientierte Einführung [online]. NWA 1.3. [Hamburg (Germany)]: Lang, H.; Schnegg, M., October 2002. Methoden der Ethnographie 1 [retrieved: 22.08.2011]. Available on World Wide Web: www.methoden-der-ethnographie.de/heft1/Netzwerkanalyse.pdf
- [35] Sethi R., Somanathan E.: Understanding reciprocity. *Journal of Economic Behavior & Organization*, 50(1): 1–27, 2003.
- [36] Simon H. A.: *The Sciences of the Artificial*. 2nd edition. Cambridge, MA: MIT Press, 1994. 247 p. ISBN 0262191938.
- [37] Smith J. H.: The Shareholder vs. Stakeholder Debate. *MIT Sloan Management Review*, 44(4): 85–90, 2003.

- [38] Sydow J.: *Management von Netzwerkorganisationen*. Beiträge aus der „Managementforschung“. 4th ed. Wiesbaden: Gabler, 2006. 472 p. ISBN 3834900974.
- [39] Tichy N. M., Tushman M. L., Fombrun C.: Social Network Analysis for Organizations. *The Academy of Management Review*, 4(4): 507–519, 1979.
- [40] Trappmann M., Hummell H. J., Sodeur W.: *Strukturanalyse sozialer Netzwerke*. Konzepte, Modelle, Methoden. 1st edition. Wiesbaden: VS Verlag für Sozialwissenschaften, 2005. 278 p. ISBN 3531143824.
- [41] Travers J., Milgram S.: An Experimental Study of the Small-World-Problem. *Sociometry*, 32(4): 425–443, 1969.
- [42] Uzzi B.: Social Structure and Competition in Interfirm Networks. The Paradox of Embeddedness. *Administrative Science Quarterly*, 42(1): 35–67, 1997.
- [43] Wasserman S., Faust K.: *Social Network Analysis*. Methods and Applications. 1st edition. Cambridge: Cambridge University Press, 1994. 825 pages. ISBN 0521382696.
- [44] Watts D. J.: *Small Worlds*. The Dynamics of Networks Between Order and Randomness. 1st edition. Princeton, NJ: Princeton University Press, 1999. 262 p. ISBN 0691005419.
- [45] Wolf C.: Egozentrierte Netzwerke. Erhebungsverfahren und Datenqualität. In Diekmann, A. *Methoden der Sozialforschung*. 1st edition. Wiesbaden: VS Verlag für Sozialwissenschaften, 2006, Kölner Zeitschrift für Soziologie und Sozialpsychologie, Special Issue 44, pp. 244–273.

Software for SNA

- [1] **RSiena**: http://www.stats.ox.ac.uk/~snijders/siena/siena_r.htm
- [2] **NetworkX**: <http://networkx.lanl.gov/>
- [3] **ORA**: <http://www.casos.cs.cmu.edu/projects/ora/software.php>
- [4] **UCInet**: <http://www.analytictech.com/ucinet/>
- [5] **Pajek**: <http://pajek.imfm.si/>
- [6] **SoNIA**: <http://www.stanford.edu/group/sonia/>
- [7] **Gephi**: <http://gephi.org/>
- [8] **Visone**: <http://www.visone.info/>
- [9] **NodeXL**: <http://nodexl.codeplex.com/>
- [10] **AutoMap**: <http://www.casos.cs.cmu.edu/projects/automap/>
- [11] **Tulip**: <http://tulip.labri.fr/TulipDrupal/>
- [12] **Vennmaker**: <http://www.vennmaker.com/>

PROGRAM GAMBIT VE VÝUCE PŘEDMĚTU ROZHODOVÁNÍ V RIZIKU A NEJISTOTĚ

THE USE OF GAMBIT SOFTWARE IN THE DECISION-MAKING UNDER RISK AND UNCERTAINTY COURSE

Pavel Stríž

Adresa: U Škol 940, Bučovice, 685 01, okr. Vyškov

E-mail: pavel@striz.cz

Abstrakt: V tomto příspěvku si představíme Open Source Software pojmenovaný Gambit, <http://www.gambit-project.org/>, který je využíván ve výuce v předmětu Manažerské rozhodování v riziku a nejistotě na Fakultě managementu a ekonomiky Univerzity Tomáše Bati ve Zlíně. Autor poukáže na ukázky z teorie her i ze základních rozhodovacích stromů.

Klíčová slova: Gambit, teorie her, rozhodovací strom, hry proti přírodě.

Abstract: In this article the author introduces Open Source Software called Gambit, downloadable at <http://www.gambit-project.org/>, which is used in the Managerial Decision-Making Under Risk and Uncertainty course at Faculty of Management and Economics, Tomas Bata University in Zlín, the Czech Republic. The author also points out at several basic game theory problems as well as a few problems which can be stated and drawn as decision tree situation.

Keywords: Gambit, Game Theory, Decision Tree, Game Against Nature.

1. Nabídka úvodem

Učiním vám nabídku! Zahráli byste si proti autorovi článku hru kámen-nůžky-papír, když možnosti volí náhodně? A co kdyby volil zarputile kámen častěji (40 %) než nůžky a papír (30 %) a svou strategii by nehodlal změnit?

2. První kroky v Gambitu

Abychom se dokázali racionálně rozhodnout, využijeme program Gambit [1], který umí řešit situace teorie her (první autorova nabídka), i rozhodovací strom (druhá nabídka), včetně zahrnutí pravděpodobnostního hráče, to je role, do které se autor příspěvku sám uvrhl.



Proč jsem ve výuce volil Gambit? Je při ruce, je to svobodný software, a navíc byla potřeba si jej stáhnout a nainstalovat. Zdánlivě triviální problém pod Microsoft Windows, nebyl všemi studenty bez kantorské pomoci vyřešen. Tudíž to bylo i zajímavé cvičení informatické.

Gambit se dá stáhnout z oficiálních stránek, www.gambit-project.org, verze z roku 2007 je obdobná verzi 2010, pro Linux, Mac OS X i Microsoft Windows. Pro Windows zvolte verzi z roku 2007, neb binárky nejsou plně připravené u verze 2010 a i při využití Cygwinu nebo MinGW bude čtenář zápasit s grafickým uživatelským prostředím (nutnost doinstalovat wxWidgets) a nemožností užití jednoho z algoritmů na 64bitových strojích. Pokud si tyto přepychové věci odpustíme, stačí po rozbalení (za předpokladu, že máte již nainstalované `gcc-c++`, `make` atd.) dát:

```
$ ./configure --disable-enumpoly --disable-gui
$ make
$ make install
```

Instalace verze 2007 na nativních Windows je přímočará. Budete mít k dispozici grafické uživatelské rozhraní i algoritmy spustitelné jako `exe` soubory z příkazové řádky. Uživatelé pracující na Linuxu či Mac OS X na tom budou z pohledu instalace lépe, většina jej nalezne v repozitáři.

2.1. Řešení problému z teorie her

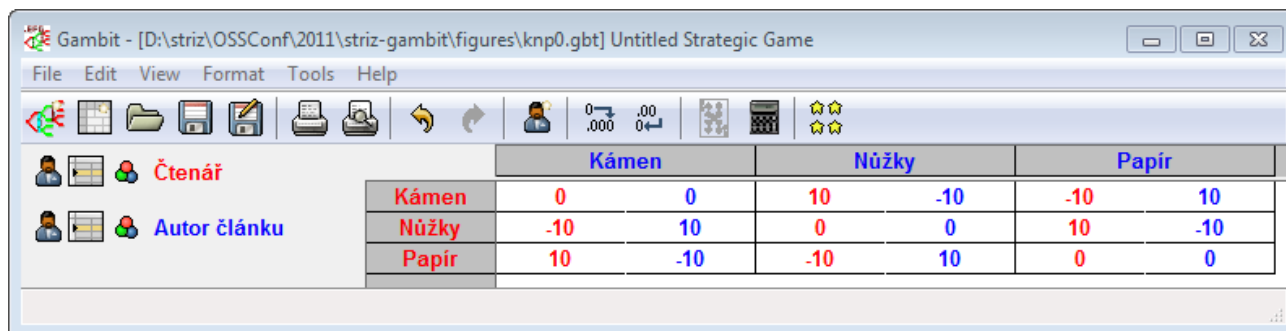
Vyřešíme si naši první situaci, což je dle klasifikace hra dvou racionálně uvažujících hráčů s nulovým součtem (co první získá, druhý ztrácí a vice versa).

Spustíme si Gambit a dáváme File→New→Strategic game. V levé části si přidáme jeden řádek a jeden sloupec díky ikonkám ,Add a strategy for this player‘. Dvojklikem změníme popisky řádků a sloupců na Kámen, Nůžky a Papír. Nyní si ohodnotíme jednotlivé možnosti her. Když vyhraji dáváme kladnou hodnotu, např. +10, když je remíza tak nula, a když prohrájeme tak -10. Červená pole (levá) jsou prvního hráče, modrá (pravá) pak druhého.

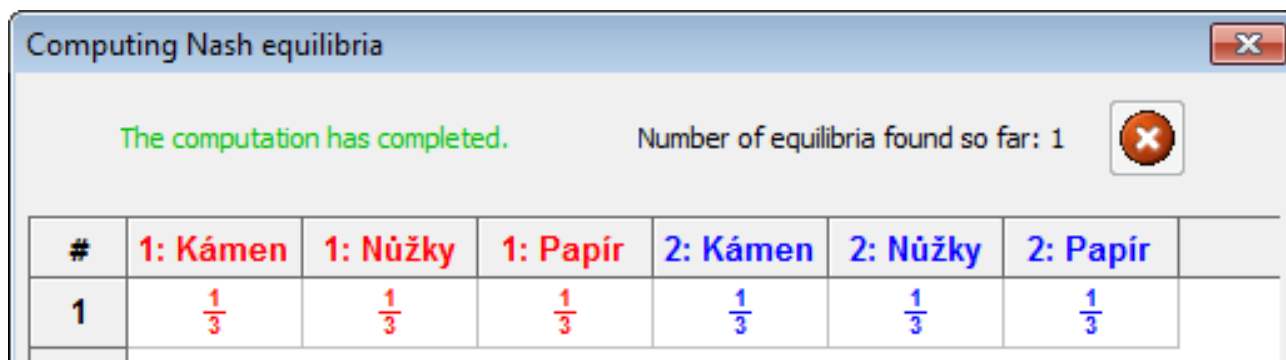
Tabulka (dvojmatice) je symetrická podle hlavní osy; v každém řádku, sloupci i dvojbuňce je součet výher nula. Dostáváme:

Po vyřešení (druhá ikonka zprava s názvem ,Compute Nash equilibria of this game‘) s předvolenými možnostmi ,Compute all Nash equilibria‘ a ,with Gambit’s recommended method‘ získáváme:

To je řešení naší hry. Každý z hráčů by měl jednotlivé možnosti (řádky, sloupce) volit náhodně a se stejnou pravděpodobností. Po stisknutí ,OK‘ v levé části pod jmény uvidíme dvakrát ,Payoff: 0‘, tedy u této hry není vítězů ani poražených v dlouhém období.



		Kámen	Nůžky	Papír
Čtenář	Kámen	0	0	10
	Nůžky	-10	10	0
	Papír	10	-10	0



#	1: Kámen	1: Nůžky	1: Papír	2: Kámen	2: Nůžky	2: Papír
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

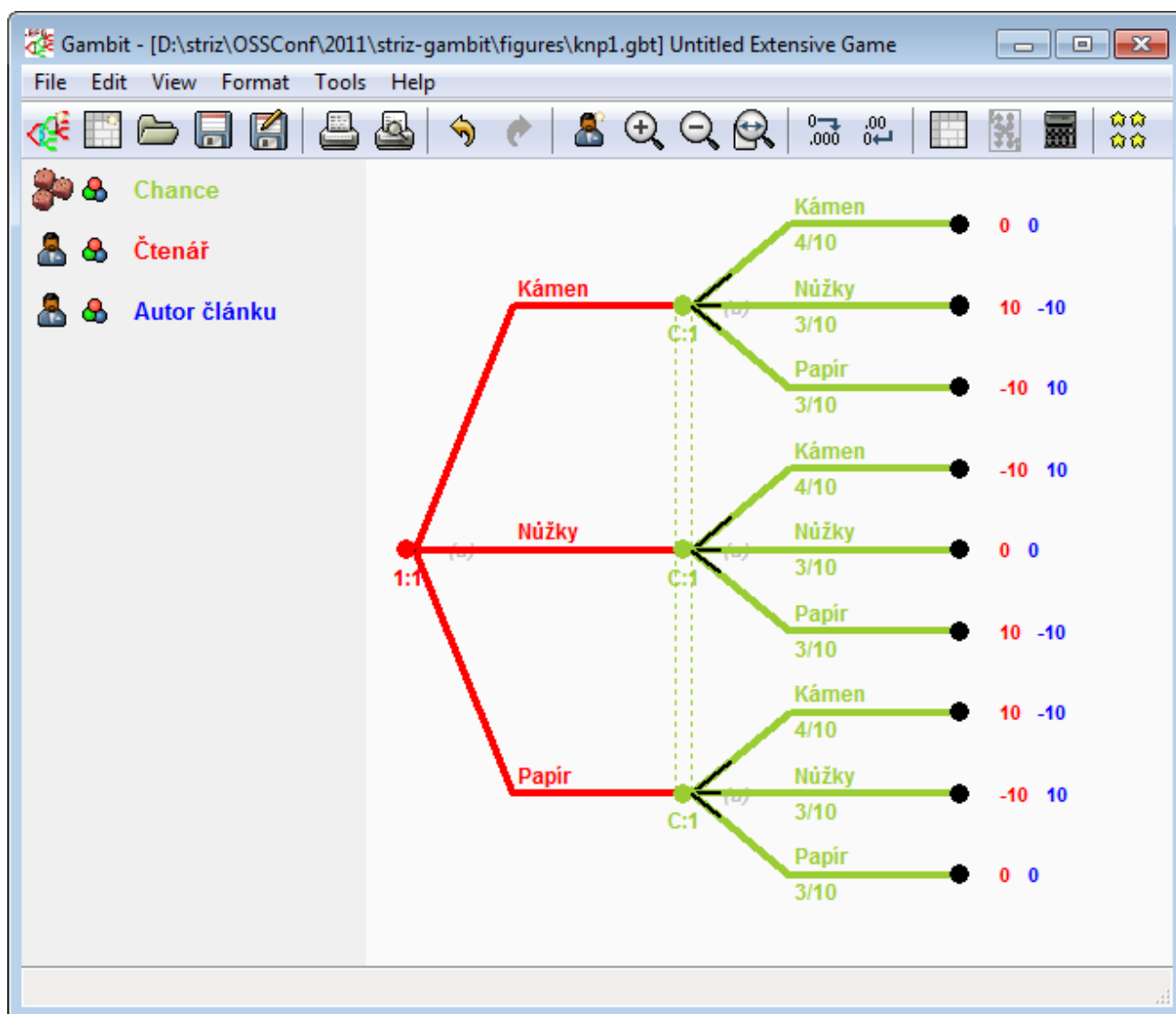
2.2. Řešení rozhodovacího stromu

Pokud se však naše volby liší a strategie je neměnná, lze hru sehrát jako hru proti přírodě. K tomu budeme potřebovat aparát rozhodovacích stromů, ten Gambit také umí.

V Gambitu si naklikneme File→New→Extensive game. Uvidíme jeden černý uzel. Na ten klikneme pravým tlačítkem myši a volíme ‚Insert move‘, změníme ‚2‘ na ‚3‘ a dáváme ‚OK‘. Objeví se nám tři nové černé uzly. Vybereme horní pravým tlačítkem na myši a volíme ‚Insert move‘. Navolíme ‚Insert move for the chance player‘ v první rolovací šipce. I zde změníme dvojku na trojku a potvrdíme ‚OK‘. Dvojklikem na červenou a zelenou ‚1‘ můžeme nastavit názvy strategií (Label) a i pravděpodobnosti (Probability) u přírody (Chance). Pole volíme dvojklikem levého tlačítka myši, volby poté potvrdíme ‚OK‘.

Pravým tlačítkem myši volíme uzel u Nůžek červeného hráče, ‚Insert move‘. Měníme na: ‚Insert move for the chance player‘, ‚at information set 1 (3 actions, 1 member node)‘, dáváme ‚OK‘. Stejně to zrealizujeme u posledního uzlu červeného hráče.

Poslední fáze je vyplnit cenové štítky, v programu Gambit zaznačeno jako ‚(u)‘ (angl. utility). Budeme jich potřebovat devět v úplně pravé části stromu. Postupně je vybereme dvojklikem a připisujeme výhry a prohry prvního a druhého hráče. Ve finále dostaneme:



Po vyřešení získáváme optimum, volit vždy Papír. Opět v levé části pod jmény hráčů uvidíme výhru/prohru. Nyní by si první hráč polepšil z nuly na jednotku výhry a druhý hráč (který se de facto hry racionálním způsobem neúčastní) ztrácí jednotku výhry:

Computing Nash equilibria

The computation has completed. Number of equilibria found so far: 1

#	1: 1	1: 2	1: 3	2: *
1	0	0	1	1

3. Pro náročnějšího čtenáře

3.1. Vězňovo dilema

Pokud je čtenář obeznám s touto disciplínou aplikované matematiky, tak jej známý problém vězňovo dilema nepřekvapí. Je to hra s nenulovým součtem a má jednu Nashovu rovnováhu (oba vězni si věří a zapírají) a jedno Paretoovo optimum (vězni si nevěří a oba druhého podvedou).

Nepřekvapí nás tedy, že mezi 140 ukázkami (podadresář /games/ ve verzi 2007; podadresář /contrib/games/ ve verzi 2010) tento problém nalezneme. Jedná se konkrétně o soubor `pd.nfg` (angl. Prisoner's Dilemma). Po otevření a vyřešení získáváme:

		Player 2	
		1	2
Player 1	1	9, 10	9, 0
	2	10, 1	0, 1

#	1: 1	1: 2	2: 1	2: 2
1	0	1	0	1

3.2. Gambit z příkazové řádky

Příjemné zjištění je fakt, že lze daný problém vyřešit z příkazové řádky, z libovolného ze tří operačních systémů. Vstupem je textový nešifrovaný soubor `pd.nfg`, který má jasně definovanou strukturu, která je popsána v manuálu.

```
$ cat pd.nfg # Linux a Mac OS X
```

Nebo zkuste následující (zmíněno bez prázdných řádků):

```
> more pd.nfg & rem Microsoft Windows
```

```
NFG 1 R "Two person Prisoner's Dilemma game" { "Player 1" "Player 2" }
{ { "1" "2" }
{ "1" "2" }
}
```

```
""
{
{ "" 9, 9 }
{ "" 10, 0 }
{ "" 0, 10 }
{ "" 1, 1 }
}
1 2 3 4
```

Řešení získáme spuštěním jednoho z algoritmů, např. na Linuxu a Mac OS X, z adresáře `/contrib/games/`:

```
$ gambit-enummixed <pd.nfg
```

Na Microsoft Windows 7 např. takto:

```
C:\Program Files (x86)\Gambit>gambit-enummixed.exe <games/pd.nfg
```

Výstupem získáváme (NE značí v angl. Nash Equilibrium):

```
Compute Nash equilibria by enumerating extreme points
Gambit version 0.2007.12.04, Copyright (C) 2006, The Gambit Project
Enumeration code based on lrslib 4.2b,
Copyright (C) 1995-2005 by David Avis (avis@cs.mcgill.ca)
This is free software, distributed under the GNU GPL
NE,0,1,0,1
```

3.3. Výhody a nevýhody

Shrneme-li, tak můžeme říci, že:

- + Program Gambit pracuje v grafickém režimu i z příkazové řádky.
- + Vstupní soubory `efg` a `nfg` jsou nešifrované, tedy snadno generovatelné.
- + Informační banner u výstupu se dá vypnout parametrem `-q`.
- + V grafickém režimu snadno získáme dominantní řádky a sloupce.
- + Verze 2007 běží zatím na všech verzích Microsoft Windows.
- + Gambit obsahuje sadu 140 ukázek, řada z nich je z odborných knih.
- + Dokumentace programu je přehledná.
- + Při uložení výsledku v grafickém režimu se stává součástí `efg` či `nfg` souboru pod tagem `analysis`.

Na druhé straně však:

- Nevýhodou zůstává, že z příkazové řádky nelze získat matici výher hráčů, ty by se musely dohledat (řešení v čisté strategii) či dopočítat (řešení ve smíšené strategii).
- Osoba neznalá teorie her může snadno narazit na situaci, kdy vybraný algoritmus zkolabuje. Pak zkolabuje i program Gambit či daný spustitelný soubor. Není to uživatelsky přívětivě řešené v grafickém režimu pod Windows a v řadě případů je potřeba ukončit proces až přes Správce úloh.
- Verze 2010 zatím neobsahuje podporu pro nativní Windows a nepodporuje 64bitovou platformu pro algoritmus `enumpoly`. Kompilace verze 2010 s grafickou podporou (wxWidgets) je pod Cygwinem na Windows 7 64bit téměř nemožná, pro neprogramátora je to nad lidské síly.
- Ačkoliv je to zmíněno v cílech autorů programu již roky, tak Gambit v současnosti nepodporuje kooperativní hry, podporuje jen nekooperativní hry a hry proti přírodě.

4. Nabídka

Závěrem opět jedna autorská nabídka ke hře s nulovým součtem, kterou kantor dává svým studentům ve výuce, a která snad zpříjemní začátek partií do rozhodovacích stromů, nám nechť poslouží na ověření si znalostí práce s Gambitem.

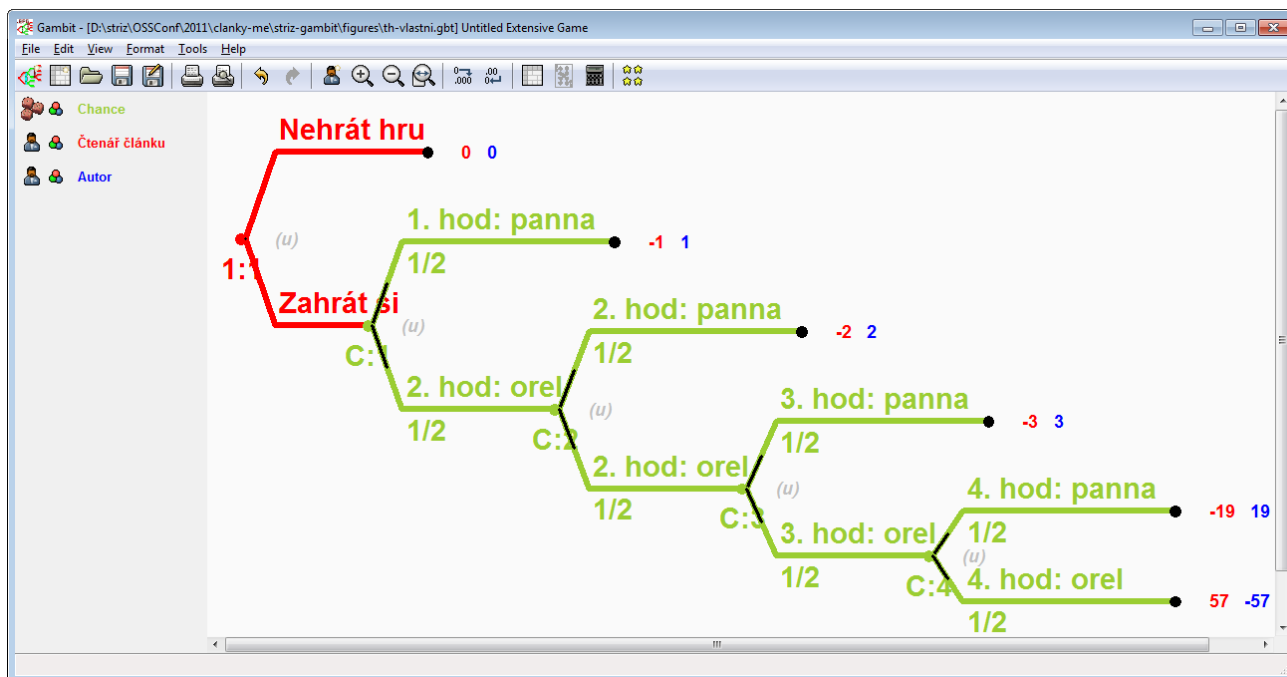
Pokud se hry nezúčastníte, nic nevyhrajete ani nic nezískáte. Pokud do hry půjdete (červený hráč), jedná se o hod korunovou mincí (z České republiky) dle následujících pravidel. Hra je smyšlená, ve výši výhry a prohry netřeba hledat hlubší smysl.

Nehrajete-li, nic nezískáte a nic neztratíte (0 Kč). Hází se mincí, maximálně čtyřikrát, pokud čtyřikrát po sobě padne orel, autor článku platí čtenáři článku 57 Kč. V opačném případě je vypláceno jemu. U prvního hodu při panně korunu a hra končí (při orlu pokračuje), u druhého dvě, u třetí tři koruny a u posledního, je-li panna, celých 19 korun.

Pokud se s autorem článku potkáte, máte možnost požadovat výhru, nebo naopak zaplatit autorovi prohru, s odkazem na tento článek. ☺

Zde je formálně zapsaná hra – hodnoty pro 2. hráče jsou postupně 0; 1; 2; 3; 19 a -57, pro prvního hráče jsou s opačným znaménkem. Abych netrpělivého čtenáře nenechal na pochybách, zmiňuji dále v článku i řešení.

S teorií her se můžeme setkat v ekonomii, politice, právních sporech [2–4], ve filmu a seriálech, kdy jeden ze zločinců si odpykává trest za svého parťáka, který ho podrazil a pak se po propuštění z vězení mstí.



Ovšem nechodme daleko, zájemce nechť shlédne film *Čistá duše – A Beautiful Mind*, 2001, který je o Johnu Nashovi, průkopníku v teorii her a nositeli Nobelovy ceny za ekonomii z roku 1994 (vedle Reinharda Seltena a Johna Harsanyie). Klíčový moment filmu a jeho netradiční řešení, které ovlivnilo celou ekonomii, je znám pod termínem bitva pohlaví (angl. Battle of Sexes nebo též Bach or Stravinsky – BoS).

Problematiku rozhodovacích stromů potkáme ve hrách jako jsou go, šachy, poker i bridž (angl. Double-dummy Problems, částečně lze použít i na Single-dummy Problems).

5. Řešení hry proti autorovi

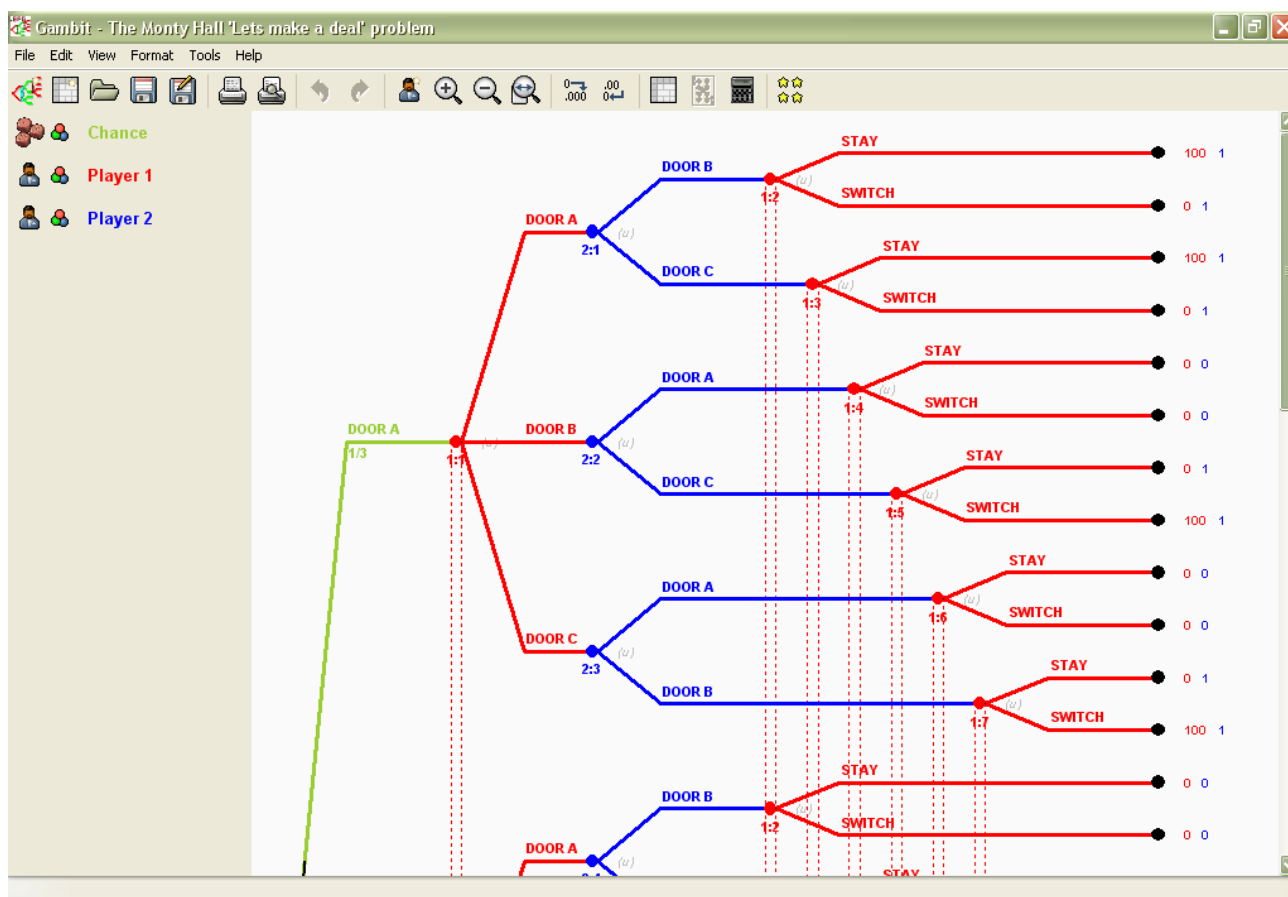
Krása úlohy zmíněné dříve spočívá v tom, že při velkém opakování takové hry je autor článku skutečně ztrátový, a to jedna koruna česká na hru. Tu si zaslouží každý čtenář, který by do hry šel, případně student, který rozhodovací strom správně ohodnotil a do hry vstoupil. V Gambitu si navíc můžeme nakliknout jednotlivé uzly, které jsou po vyřešení numericky ohodnoceny.

V reálném životě to však znamená, že autor článku by takovou hru hrál s každým zájemcem jen jednou a je tu sportovní šance, že čtyřikrát po sobě autorovi článku orel nepadne a na svých ztrátových 57 Kč si pěknou řádku her počká, tedy většina čtenářů by naopak musela autorovi platit, i když jen položku korunovou. Autor článku je vždy ochotný uznat svou prohru v dlouhodobém horizontu a onu korunu každému účastníkovi/studentovi uhradit

bez ohledu na výsledek hodů, ať už z pohledu sportovního nebo kantorského. A kdo nehraje, nevyhraje, ani nemá šanci přijít si ke ztrátě!

6. Závěr

Gambit lze použít i na komplexnější situace, bez hlubšího rozboru naznačím jedním snímkem obrazovky (`montyhal.efg`), kterým se zároveň rozloučím (Monty Hallův paradox alias problém tří dveří, viz též film *Oko bere*, 2008).



Literatura

- [1] McKelvey R. D., McLennan A. M., Turocy T. L.: *Gambit: Software Tools for Game Theory*, Version 0.2010.09.01, 2010. Dostupné na serveru: <http://www.gambit-project.org/>
- [2] Chobot M., Turnovcová A.: *Modely rozhodovania v konfliktných situáciách a za neurčitosti*, Alfa, Bratislava, 1980.
- [3] Mañas M.: *Teorie her a její aplikace: Učebnice pro studenty VŠE*, 1. vyd. Praha: Státní nakladatelství technické literatury, 1991. 278 s.
- [4] Pitel, J. a kol.: *Ekonomicko-matematické metody*. Bratislava, 1988.

NĚKOLIK POZNÁMEK K VÝVOJI HOSPODÁŘSKÉ STATISTIKY V ČR

Jaroslav Češka

Adresa: Přístoupimská 12/394, 108 00 Praha 10

Abstrakt: Transformace československé ekonomiky po roce 1989 spojená s dalšími významnými změnami (zánik ČSFR, vstup ČR do EU) vedly k významným změnám v příslušných úsecích ekonomické statistiky. Na úseku makroekonomických ukazatelů byl zaveden systém národního účetnictví, který nahradil dříve používaný systém bilancí národního hospodářství. Úpravy uskutečněné v hospodářské statistice ČR po vstupu do EU zvýšily prostorovou srovnatelnost zkoumaných ukazatelů. Záměna národních ukazatelů novými však ztížila na některých úsecích dlouhodobou časovou srovnatelnost.

Klíčová slova: Transformace ekonomiky, ekonomická statistika, ukazatele.

Abstract: Transformation of the Czechoslovak economy after the year 1989 together with the ČSFR disintegration in 1992 and the accession of the ČR to the EU in 2004 caused great changes in the corresponding parts of economic statistics. In the area of macroeconomic aggregates System of National Accounts was introduced replacing previously used System of National economy balances. An adaptation of the Czech economic statistics to the new situation enlarged meaningful of the Czech indicators with those of other countries. Replacement of national indicators by the new ones, however, caused difficulties in the long-term comparisons.

Keywords: Economic Transformation, Economic Statistics, Indicators.

V Informačním bulletinu České statistické společnosti (ročník 22, číslo 4, 4. prosince 2010) byl zveřejněn příspěvek Stanislavy Hronové a Richarda Hindlse „Hospodářská statistika z pohledu 20 let vývoje“, který je velmi zajímavý a poučný. Je zaměřen na problematiku makroekonomických agregátů, zejména na zavedení systému národního účetnictví (SNA) a v menší míře na některé související statistické úseky (ceny, zaměstnanost aj.).

Příspěvek ve svém výkladu o hospodářské statistice, definici ukazatelů, vztahu k ekonomické teorii, srovnatelnosti aj. je vhodné, podle mého názoru, doplnit. Některá tvrzení jsou totiž až příliš „kategoricky“ formulována, případně nabízí zevšeobecnění z poznatků, které mají opodstatnění jen v určité části ekonomické statistiky. Níže uvedené poznámky jsou omezeny na poznatky z oficiální statistiky a nekladou si nárok na úplnost.

Hospodářská statistika ve svém souhrnu i na řadě svých úseků se vyvíjela nezávisle na systému bilancí národního hospodářství (MPS) nebo systému

národního účetnictví (SNA). Časově i logicky předchází tvorbě makroekonomických ukazatelů a jejich systémů. Vazba ukazatelů ekonomické statistiky na soubory ukazatelů MPS a SNA je v některých případech nepřímá, částečná nebo i značně zprostředkovaná. Nejde přitom jen o omezenost vyplývající z měrných jednotek používaných v uváděných systémech.

Vztah mezi ekonomickou teorií a ekonomickou statistikou je, jako v případech jiných disciplín, velmi mnohotvárný, vzájemně se obohacující. Není, podle mého názoru, jednostranná posloupnost, napřed teorie a pak statistika, máme-li na mysli ekonomickou statistiku jako praktickou činnost.

V případě údajů a ukazatelů ekonomické statistiky je nutno přiznat, že ne vždy je sama definuje. Ekonomická statistika se ve své gnoseologické činnosti opírala a opírá i o vymezení a definice jiných vědních oborů, vymezení v právních a účetních normách, technických i jiných disciplínách. Vzhledem k potřebě srovnatelnosti navazuje i na metodiku použitou v předchozích statistických zjišťováních.

V některých konkrétních případech, zejména na nových úsecích ekonomické statistiky, oficiálních statistika pracovala bez těchto opor a musela tak sama příslušná vymezení určovat (soubory, statistické a zpravodajské jednotky, zjišťované znaky apod.). Příkladem bylo vymezení nových výrobků při jejich zjišťování v průmyslu, vymezení výzkumné a vývojové základny apod. Klasickým příkladem je statistika rodinných účtů, obsah (program) mikrocensů i jiných výběrových šetření u domácností.

Řada odvětvových a průřezových ekonomických statistik se zjišťuje a publikuje v přirozených (naturálních) měrných jednotkách. Vzhledem k členství ČR (dříve ČSFR) jsou ukazatele a sestavy z těchto úseků předkládány příslušným mezinárodním vládním organizacím (Statistickému úřadu OSN, EKH OSN, ILO, FAO, UNESCO, WHO aj.), které je uvádějí ve svých statistických i jiných publikacích. Metodika požadovaných ukazatelů byla vypracována skupinami statistických expertů a schválena odpovídajícími orgány (např. výbory Evropské hospodářské komise), které také rozhodovaly o jejich svolání. V rámci Mezinárodní organizace práce (ILO) byla metodika požadovaných ukazatelů projednávána na konferencích statistiků práce. Vzhledem k tomu jsou údaje zemí ve statistických publikacích těchto organizací vzájemně srovnatelné.

Obdobně bylo postupováno při přípravě jednorázových statistických publikací Evropské hospodářské komise (EHK)¹ i při přípravě evropských variant světových doporučení k provádění censů v příslušných úsecích hospodářských

¹Např. z více než 10 pravidelných statistických publikací EHK, jen u jedné (Short-term indicators) byly publikované údaje za jednotlivé země ponechány na národních metodikách.

statistik (v průmyslu, v zemědělství, při sčítání lidu, domů a bytů aj.). I obsah těchto zjišťování, včetně publikování jejich výsledků, byl projednáván na zasedáních statistických expertů svolávaných mezinárodními statistickými orgány. Českoslovenští statističtí odborníci se popsanych metodických unifikačních prací aktivně zúčastňovali.

Metodické unifikační statistické práce byly rovněž řešeny pod patronací a v rámci sub-regionálních statistických orgánů (Statistický úřad EHS, Stálá statistická komise RVHP), jejichž zástupci se zúčastňovali jednání Konference evropských statistiků i dalších orgánů EHK OSN.

Zvláštní zmínku si zaslouží i srovnávací práce na systémech MPS a SNA uskutečňované pod patronací Statistické komise OSN a KES v dřívějších letech, včetně jednání příslušné skupiny statistických expertů, kterých se zúčastňovali pracovníci FSÚ.

Jak správně uvádějí autoři příspěvku, i v hospodářské statistice ČR byla od 1. ledna 1993 uplatněna řada významných změn. Hospodářská statistika se i v průběhu dalších let vyvíjela a na některých úsecích také měnila.

Na úseku vnějších faktorů určitý mezník představuje vstup ČR do Evropské Unie 1. května 2004. Naše statistická služba se tak stala součástí Evropského statistického systému, včetně všech povinností v zajišťování a předkládání požadovaných statistických údajů a ukazatelů ve stanovené metodice (podle schválených manuálů) a lhůtách. Požadavky EU mají právní oporu v nařízeních orgánů EU (Rady nebo Komise EU). Harmonizací československé statistiky v souladu s požadavky EU byly vytvořeny potřebné podmínky pro prostorové srovnávání údajů ČR s údaji dalších členských zemí EU pro potřeby uvedených orgánů.

Harmonizace národních statistik v rámci EU realizovaná v praktické statistické činnosti ČR i záměnou dříve používaných statistických zjišťování novou metodikou šetření, včetně definic ukazatelů, ztížily však na některých úsecích hospodářské statistiky ČR (ČSFR) časovou i věcnou dlouhodobou srovnatelnost navazujících statistických údajů a ukazatelů.

Vyjdeme-li z účelu statistických zkoumání ekonomických jevů a procesů v ekonomice ČR, pak „zásadní“ změnu hospodářské statistiky nelze vysvětlit jen jako důsledek záměny MPS systémem SNA. Ze srovnání souboru publikovaných ekonomických statistik například v ročenkách ČSFR a ČR lze vyvodit, že změny (mimo kapitoly makroekonomických ukazatelů) byly vyvolány zejména potřebou:

a) zjišťování nových skutečností (příklad: soubory ukazatelů charakterizující nezaměstnanost),

b) zjišťování a publikování ukazatelů, které při dřívějším pojetí byly sledovány a publikovány jen na úrovni federace (příklad: statistika zahraničního obchodu, dopravy apod.),

c) zjišťování a publikování dříve neveřejných údajů (příklad: vojenské výdaje, platební bilance, devizové rezervy, zadluženost),

d) prohloubení statistických zjišťování o ukazatele soukromého sektoru, ukazatele podle nových klasifikací, třídících znaků a sektorů, ukazatele o výrobních a jiných jednotkách pod zahraniční kontrolou,

e) zjišťování a publikování statistik tříděných podle vlastnických forem,

f) zjišťování k zajištění požadavků a potřeb EU i dalších mez. organizací,

g) náhradou vyčerpávajících zjišťování výběrovými šetřeními, uplatněním finančních hledisek místo věcných znaků apod.

V některých případech došlo k omezení statistických zjišťování, některé ukazatele se přestaly v důsledku nových právních úprav a jiných příčin sledovat a publikovat. Změny se realizovaly nejen v souborech statistických a zpravodajských jednotek, ale i v souborech zjišťovaných znaků. Některá statistická zjišťování byla také po určité době obnovována, jako např. zjišťování údajů o přírůstcích a úbytcích bytů podle příčin apod.

Harmonickému rozvíjení jednotlivých úseků hospodářské statistiky neprospívá jednosměrný tlak na uzpůsobení odvětvových a dalších dílčích statistik potřebě sestavování souhrnných agregátů. Ten vyvolává omezování v zjišťování statistických údajů a ukazatelů v přirozených (naturálních) jednotkách, včetně jejich sestav, což se nepříznivě projevuje v redukci poznání a přínosu oficiální statistiky na tomto úseku.

K neúplnosti a omezené obsahové srovnatelnosti časových řad ukazatelů za dříve existující územní celky přispělo i souběžné uplatnění změn, včetně klasifikací ještě v posledních letech existence ČSFR. Významné změny v souborech zpravodajských jednotek i zpravodajských sítích, velikostní a jiná neustálená omezení ve zpravodajských povinnostech vedoucí k náhradě dříve používaných vyčerpávajících zjišťování vedly ve svém souhrnu ke složitějším podmínkám pro oficiální statistiku na úseku statistických zjišťování. Nesporný vliv měly i posuny v souboru hlavních uživatelů, zaměření a způsoby jejich činnosti projevující se ve změnách požadavků na oficiální statistiku v hospodářské oblasti.

Zvláštní pojednání by si zasloužil i popis vývoje československé ekonomické statistiky po 2. světové válce, o kterém se autoři článku zmiňují. I když výklad je zaměřen na souhrnné agregátní ukazatele, zevšeobecné tvrzení o odlišnostech a nesrovnatelnosti celých systémů používaných ukazatelů v zemích i jejich publikování je zavádějící a vyžadovalo by to hlubší vysvětlení.

Československá statistika na ekonomickém i dalších úsecích musela totiž po 2. sv. válce řešit řadu naléhavých a konkrétních úkolů, jako soupisy škod v jednotlivých odvětvích NH a u obyvatelstva, obnovit statistická zjišťování na celém předválečném území ČSR, obnovu zpravodajských sítí na dříve odstoupeném území, překonat rozdílný vývoj oficiálních statistik na Slovensku a bývalém Protektorátu při předkládání údajů za ČSR mezinárodním organizacím, zabezpečovat agendu spojenou s vydáváním osvědčení v národní příslušnosti aj.

POZNÁMKY Z KONFERENCÍ ROBUST A COMPSTAT 2012

Nikola Kaspříková

E-mail: data@tulipany.cz

1. Úvod

V roce 2012 proběhly statistické konference ROBUST a COMPSTAT. Spousta informací o obsahu konferencí je k nalezení na webových stránkách jim věnovaných, zkusím se zaměřit na ten zbytek.

2. ROBUST

Ve dnech 9. – 14. září 2012 se v Němčičkách konala konference ROBUST. Se jménem Němčičky se zřejmě valná většina účastníků konference seznámila až v souvislosti s ROBUSTem. Nicméně dobře známou oblast, ve které najdeme nedaleko Němčiček taky třeba Velké Pavlovice, Bořetice, Kobylí nebo Čejkovice, asi ani není nutné čtenářům představovat.¹

Při pokusu o shlukovou analýzu přednesených příspěvků (provedenou osvědčeným algoritmem Pohled z okna, který obvykle vykazuje vynikající vlastnosti jak z hlediska (jakýmkoli způsobem měřené) kvality výsledné klasifikace, tak z hlediska výpočetní náročnosti) se jako nejvíce zastoupené shluky jevíly analýza kompozičních dat, aplikace ve financích a pojišťovníctví, hloubka dat nebo řízení jakosti. Určité oživení v sále vyvolala zajímavá prezentace vedená v barvách jednoho ze sponzorů soutěže studentů a zároveň

¹Je to přece jasné – z Hodonínska například pocházel první československý prezident a v Čejkovicích chodil několik let do školy.

jednoho z významných dodavatelů řešení² pro analýzu dat a podporu rozhodování. Vystoupení bylo věnované diskuzi praktické aplikace metod počítačové lingvistiky a text miningu. Byl to jeden z příspěvků, které vyvolaly docela výraznou odezvu, dal podněty k zamyšlení o zaměření aplikací (určitě nejen v diskutovaném případě) a poskytl příležitost si zase jednou připomenout, že my lidé někdy o něco usilujeme a k prosazování svých zájmů používáme taky analýzu dat.

3. COMPSTAT

Konference se konala koncem srpna v kyperském Limassolu, jen pár stovek metrů od archeologického naleziště Amathus. Účastníci konference měli možnost seznámit se i s dalším archeologickým nalezištěm Kourion – tam se totiž konal slavnostní koncert „Towards Europe of Peace“ uspořádaný při příležitosti právě probíhajícího kyperského předsednictví Radě EU. Koncert se konal v původním starověkém amfiteatru, takže jsme mimo jiné mohli zjistit, jaké to je posedět si v klidu dvě hodinky na kameni. Vedle dvou skladeb u nás spíše méně známých autorů zazněla Beethovenova devátá symfonie³. V prostředí, ve kterém se sešlo několik stovek vědců a které možná mělo připomínat ducha evropské pospolitosti, se mi při poslechu Beethovena bohužel vybavila akorát jedna knížka Anthonyho Burgesse (a film S. Kubricka). Anthony Burgess se později připomněl ještě jednou, když některým cestovatelům zkomplikovala návrat z konference stávka zaměstnanců jedné letecké společnosti.⁴ Zmíněná symfonie má čtyři věty a zajímavé bylo zjištění, že mezi statistiky je celkem dost populární potlesk mezi větami. Z toho pohledu Stanley Azen, kterému se v hudbě dostalo i obстоjného formálního vzdělání, udělal dobře, když hymnu IASC (která během jeho vystoupení na COMPSTATu taky zazněla) navrhl jako poměrně nekomplikovaný kus. V matematice člověk nikdy neví, jak to dopadne s tou příští větou, takže raději hned vděčně zatleskat za každou, která se už povedla. Ale na koncertě uděláme líp, když vyčkáme, až dirigent přestane být zády.

Na ostrově všudypřítomné kočky se ukázaly i během koncertu, jedna dokonce napochodovala přímo mezi hudebníky a zasloužila se tím o probuzení některých jinak už bezmála podřimujících kolegů. O příčinách zjevného bla-

²Prodávat software už by dneska nestačilo, dodává se taky znalost business procesů, případně rovnou i tým konzultantů, od kterých se očekává, že problém vyřeší.

³Její část je hymnou EU.

⁴Název knihy popisující kromě dalšího také vizi řádění syndikalistů vypočteme v softwaru R [5] příkazem `sum(x, 1)`, kde za `x` dosadíme název románu George Orwella (vybíráme titul vhodného datového typu).

hobytu koček na ostrově možná něco napoví projev psa v románu [4], který praví „Podle mě celá ta záležitost souvisí s tím, jak si náš Prorok Muhammad raději uřízl kus suknice, než aby probudil kočku, která na ni podřimovala.“ Řeč je pochopitelně o kočce čtyřnohé a (mohlo by to zajímat chovatele koček) vhodným řetězcem pro zahájení rozhovoru s webovým vyhledávačem je v tomto případě *Muezza*. Nutno poznamenat, že Limassol se nachází v řecké, nikoli v turecké části ostrova, ale kočky se zřejmě politickými hranicemi tolik netrápí. Jinde je ale rozdělení ostrova patrné velmi jasně a má četné projevy, které vnímají i turisté – například speciální známku *refugee stamp*, která se povinně přidává na poštovní zásilky a je zvláštním poplatkem určeným na pomoc lidem, které postihly události z roku 1974.

Hodně prostoru na konferenci dostal Stanley Azen⁵. Převzal za svou práci ocenění od zástupce IASC a (jako šéfredaktor časopisu CSDA) od zástupce společnosti Elsevier. Ve svém vystoupení nejprve připomněl významné momenty své kariéry v biostatistice, hudbě a výpočetní statistice. Stanley Azen před lety napsal spolu se svým učitelem Aldelmonem Afim⁶ učebnici [1] a teď zdůraznil nezastupitelnou roli ještě dalšího, dříve neuváděného, spoluautora⁷. Nakonec se zmínil o některých projektech, na kterých se v poslední době podílel (mimo jiné šlo o studii zaměřenou na zlepšení životního stylu seniorů), a upozornil na prospěšnost mezioborových spoluprací v biomedicíně výzkumu.

Příští COMPSTAT by měl být v roce 2014 v Ženevě. Hlavním organizátorem bude Manfred Gilli⁸, který v Limassolu vedl pěkný tutoriál v duchu knížky [3], ve kterém mimo jiné vyjádřil přesvědčení, že často je v aplikacích rozumné zvolit raději heuristické metody a modely, které lépe odpovídají řešenému problému, než počítat přesně (vzhledem k povaze úlohy třeba i zbytečně moc) v modelu, který vznikl zavedením hrubě zjednodušujících předpokladů.

4. Závěr

Snad každá konference má nějaké zvláštní, neopakovatelné vlastnosti, díky kterým si ji účastníci budou ještě dlouho pamatovat.

Nicméně konference mají taky mnohé společné prvky. Za pozornost určitě stojí přímý přenos know-how mezi konferencemi prostřednictvím fenoménu znovupoužitých (a postupem času jako víno zrajících) prezentací. Hezky to lze

⁵Který má web – <http://www.stanazen.com/>.

⁶Viz <http://www.biostat.ucla.edu/Directory/afifi>.

⁷Spoluautor má taky web – <http://www.jackdaniels.com/>.

⁸Viz <http://www.unige.ch/ses/dsec/static/gilli>.

vyjádřit slovy „I’m giving the same presentation“. Takovou větu sdělil trochu smutně jeden z významných (posuzováno mj. podle prezentaci přiděleného času) účastníků COMPSTATu svému známému, se kterým se zřejmě potkal i na nějaké předchozí konferenci. Je třeba v této souvislosti taky poznamenat, že některé prezentace, zejména tutoriály, byly zjevně naplánovány na úplně jinou délku než na čas přidělený na COMPSTATu. Někteří přednášející v dokumentu s prezentací rádi připomínají auditoriu, kde a při jaké příležitosti se zrovna sešli. Posluchači se pak s využitím takové nápovědy třeba v Limassolu dozvědí, že jsou zrovna v Avignonu a o tři měsíce mladší, nebo v Němčičkách zjistí, že si mohou jít zaplavat na Long Beach (ale je pravda, že bazének na ROBUSTu byl). A v tu chvíli jsou pseudoreplikace skutečným problémem. Ale možná by se v tom dal rozpoznat projev konceptu *Reproducible statistical research*, v tomto případě přesněji *Reproducible statistical presentation*, což by naopak mohlo být žádoucí. Navíc si myslím, že přiměřená recyklace, vzhledem k poměrně vysoké hustotě konferencí možná jediný zvládnutelný přístup, má i spoustu výhod a především je to postup v podstatě ekologický, takže správný.

Podstata konferencí je již docela důkladně prodiskutována v literatuře a je těžké k tomu dodat ještě něco nového. S přehledem se uměl vyjádřit například autor, který se později za odměnu stal nositelem Řádu britského impéria – kdo snad ještě nebo už nezná, může alespoň letmo zabrousit do kroniky ROBUSTů z minulého tisíciletí a přečíst si příslušné odstavce, které jsou citované v dodatku článku [2]. Osudy postav románů Kingsleyho Amise, dalšího z autorů britského univerzitního románu, radši (zejména v souvislostech s ROBUSTem) podrobněji rozebírat nebudu, aby si to náhodou někdo z pánů kolegů nevyložil nějak zle.

Literatura

- [1] Afifi A. A., Azen S. P.: *Statistical analysis: a computer oriented approach*. Academic Press, 1972.
- [2] Coufal J.: *Abstraktní pojem čísla a měření v historii aneb máte u někoho vroubek?* ROBUST 96. Sborník prací 9. letní školy JČMF. Jednota českých matematiků a fyziků, 1996.
- [3] Gilli M., Maringer D., Schumann E.: *Numerical Methods and Optimization in Finance*. Academic Press, 2011.
- [4] Pamuk O.: *Jmenuji se červená*. Argo, 2007.
- [5] R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0, URL: <http://www.R-project.org/>.

Obsah

Vědecké a odborné statě

Zdeněk Wagner, Miroslav Fridrich, Jaroslav Schwarz, Vladimír Ždímal
Zpracování hromadných dat distribuce velikosti aerosolových částic v reálném čase a jejich prezentace v evropských i celosvětových databázích 1

Eva Eckenhofer
Application of Social Network Analysis in Business and Science 14

Jiné statě

Pavel Stríž
Program Gambit ve výuce předmětu Rozhodování v riziku a nejistotě 40

Jaroslav Česka
Několik poznámek k vývoji hospodářské statistiky v ČR 49

Zprávy a informace

Nikola Kaspříková
Poznámky z konferencí ROBUST a COMPSTAT 2012 53

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Sokolovská 83, 186 00 Praha 8. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in bulletin are published in English, Czech and Slovak languages.

Předsedkyně společnosti: prof. Ing. Hana ŘEZANKOVÁ, CSc., KSTP FIS VŠE v Praze, nám. W. Churchilla 4, 130 67 Praha 3, e-mail: hana.rezankova@vse.cz.

Redakce: prof. Ing. Václav ČERMÁK, DrSc. (předseda), prof. RNDr. Jaromír ANTOCH, CSc., prof. RNDr. Gejza DOHNAL, CSc., doc. Ing. Jozef CHAJDIK, CSc., doc. RNDr. Zdeněk KARPÍŠEK, CSc., RNDr. Marek MALÝ, CSc., doc. RNDr. Jiří MICHÁLEK, CSc., prof. Ing. Jiří MILITKÝ, CSc., doc. Ing. Josef TVRDÍK, CSc., Mgr. Ondřej VENCÁLEK, Ph.D.

Redaktor časopisu: Mgr. Ondřej VENCÁLEK, Ph.D., ondrej.vencalek@upol.cz.
Informace pro autory jsou na stránkách společnosti, <http://www.statspol.cz/>.

DOI: 10.5300/IB, <http://dx.doi.org/10.5300/IB>
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.