Informační soukromí – pohledy

Vašek Matyáš

PV080

Úroveň ochrany osobních dat

- *Rozhodujícím ukazatelem úrovně ochrany je cena osobních dat „na ulici“ – na černém či šedém trhu*. (Roger Needham, Cambridge U.)
  - Zdravotní data „běžné“ osoby v UK lze získat za cca 150-200 liber
  - V kanadské provincii Quebec podle některých „inzerátů“ 20-60 liber.
  - Podle Needhama by měla cena být výrazně nad 500 liber.

Příklad cen v UK (2006)

Požadovaná informace	**Platba komplicům**	**Cena-klient**
Adresa	neznámá	£17.50
Adresa podle tel. čísla	£40	£75
Adresa podle tel. čísla (mobil)	neznámá	£75
Seznam členů rodiny a přátel	£60 – £80	neznámá
Údaje o vozidle z registru	£70	£150-200
Trestní rejstřík	neznámá	£500
Tel. číslo – blokované	£40	£65 – £75
Výpis z účtu mobilního telef.	neznámá	£750
Údaje z řidičského průkazu	neznámá	£250

Mikrostudie studentů FI 2017

- Inzerát s několika gramatickými chybami (pro podpoření pozadí identity) a poptávka po DB pro prodej, s požadavkem na fyzické adresy osob, případně na jiné citlivější údaje.
- Záznam *s fyzickou adresou, telefonem a emailem cca 15 haléřů / subjekt*.
- Záznam *s rodným číslem nebo číslem občanského průkazu 5,17 Kč / subjekt.*

Cenu osobních dat ovlivňují

1. Výše trestu těm, kdo data jiných řádně neohlídali a spolupodíleli se tak na jejich úniku.
2. Výše trestu těm, kdo s nimi neoprávněně manipulují.
3. Úroveň ochranných mechanismů.

Postoj občanů k zacházení s osobními daty (Anglie, 90. léta)

- Necelých 20 % občanů totálně lhostejných,
- Stejný počet velmi obezřetných až paranoidních
- Asi 60 % je ochotno část svých práv nechat omezit za “přiměřenou úhradu” - finanční, věcnou či nejčastěji v podobě výrazného zlepšení služeb.

Průzkum v Německu – I.

- *Privacy in e-commerce: stated preferences vs. actual behavior (Berendt a kol.), ACM Communications, April 2005*
- Soukromí si chránící – 30 %
- (Téměř) lhostejní – 24 %
- Citliví na profilování – 26 %
- Citliví na identitu – 20 %

Průzkum v Německu – II.

- Za určitých okolností je ovšem většina uživatelů online ochotna zapomenout na zábrany a sdělit osobní informace i bez skutečně závažných důvodů (takto učinit)
- I uživatelé, kteří podle vlastního názoru jsou citliví na ochranu osobních dat, tak při online interakci nekontrolují v tomto směru své chování

Experiment v Cambridge

- *How Much is Location Privacy Worth? (Danezis a kol.)*
- Info studentům 1. ročníku o placeném výzkumu se sběrem informací o jejich pohybu (mobil – 28 dnů, 24 hodin denně)
  - Aukce!!!
- £10 medián, £27.4 průměr (max. £400, min. 0)
- Se zvážením prodeje pro komerční účely pak £20 medián, £32.8 průměr (max. £300, min. 0)

Obdobný experiment ve větším měřítku...

- Následující slajdy jsou výjimečně v angličtině ☺
  - Prezentace připravená v souvislosti s rozborem výsledků studie...

Starting Points

- Privacy – ensured by legal system or by technology
- Technologies to preserve privacy are really expensive
- Yet privacy intrusive technologies become more common
  - GSM system used for tracking down particular handsets (more precise than needed for the GSM system itself)
- What is the value of privacy?
  - How much are people willing to pay to protect their privacy (location privacy in this case)
  - What are black market prices and penalties
    - UK: £17.50 for address; up to £500 for criminal records check; £750 for mobile phone account details. (UK IPC, May 2006)
    - UK – penalties for privacy breaches in low £'000 per individual
    - US health data (HIPAA) – civil penalty $100 per violation
- Design a study about how much we want to get for being tracked 24/7

New Study

- Organised within FIDIS project (www.fidis.net)
  - Spring 2006
  - Pseudonymity, with only email address provided
- Five countries involved
  - Belgium, Czech Republic, Germany, Greece, Slovakia
- Information advertised to
  - University students (IT) – all countries
  - University students (regardless on study) – CZ, DE, SK
  - Mobile phone community – CZ, DE

atalog mobilů
berte značku -
berte model -
nání mobilů
ání podle parametrů

Mobilní Big Brother: nechte se sledovat – vědecky!

7. 3. 2006, Marek Lutonský

formát pro tisk

Bezpečnostní laboratoř **BUSLab** brněnské Masarykovy univerzity a Vysokého učení technického se chystá uspořádat průzkum, který do jisté míry připomíná reality show typu Big Brother. V rámci třicetidenního průzkumu se prostřednictvím mobilního telefonu bude pravidelně zjišťovat poloha přihlášených účastníků.

Dostal se k nám text oběžníku, který koluje po vysokých školách:

> Potrebujeme ke spolupraci jednotlivce, kteri budou timto zpusobem monitorovani pro ucely sociologicko-technologicke studie o mobilite lidi.
>
> Hlavni motivaci je pak vhodnost existujicich struktur siti mobilnich telefonu s ohledem na potreby uzivatelu mobilnich telefonu. Upozornujeme, ze v prubehu experimentu by ucastnici nemeli vypinat sve mobilni telefony.
>
> Nasbirana data o poloze budou uchovana a mohou byt v budoucnu vyuzita pro dalsi akademicky vyzkum. Poloha mobilniho telefonu bude zjistovana kazdych 5 minut, a to 24 hodin denne, 7 dni v týdnu ve spolupraci s mobilnim operatorem. Tato cinnost bude provadena po celou dobu experimentu, tj. 30 dni. Presnost zjistovani polohy zavisi na tom, ke kolika "bunkam" (vysilace mobilnich operatoru) je telefon v danou dobu prihlasen - zhruba 800 metru ve slabe osidlenych mistech (venkov) a 100-200 metru v

Organisation

- First form (webpage)
  - Language
  - Background (computers, law, other)
  - Gender
  - Network operator used (list of local operators)
  - Do you carry your mobile all the time?
  - How often are irregular movements (hourly, daily, weekly, monthly)?
  - Who do you talk to (friends, family, partner, business)?
- Second form
  - Commercial exploitation (decline, same bid, revised bid)
- Third form
  - Commercial use for one year (decline, write the bid)

Demographics

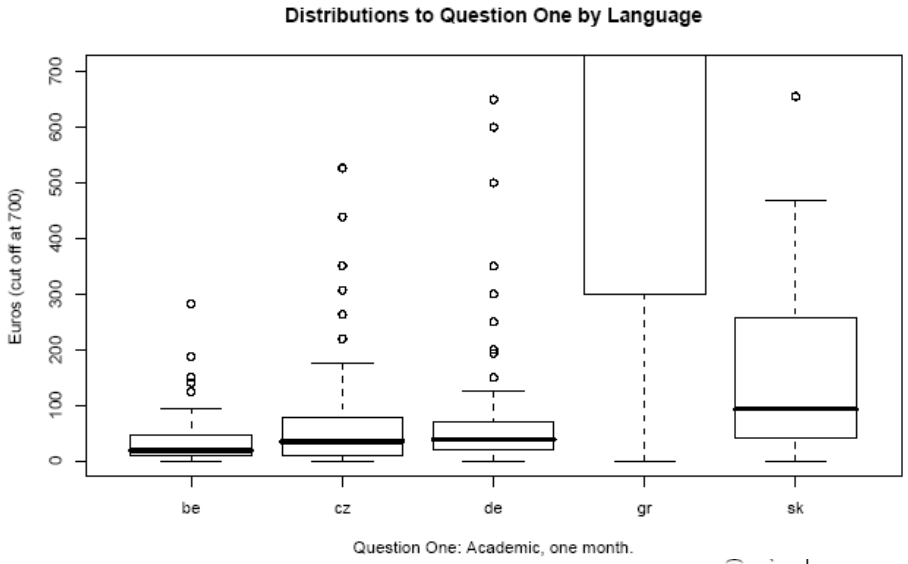
- Number of participants per country
  - Belgium 37/3 (no of participants/females)
  - Czech Republic 744/131
  - Germany 251/33
  - Greece 30/6
  - Slovak Republic 152/46
- Students in all countries, mobile phone communities in Czech Republic and Germany
- Size of sample sets
  - Czech Republic, Germany, Slovak Republic – deep analyses
  - Belgium, Greece – too small, control sets

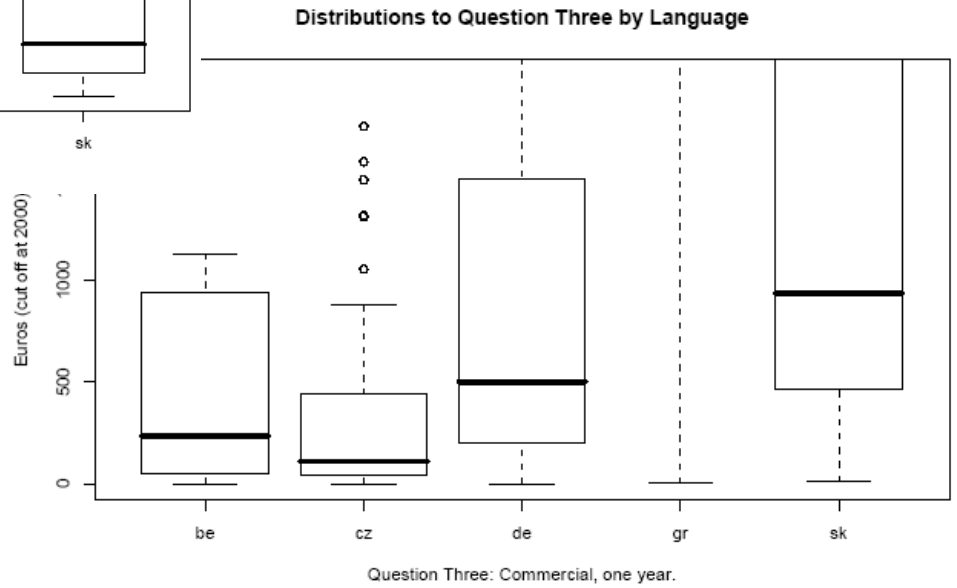
Cautiousness

- Drop-out rates
  - Early drop-outs (239 out of 2582)
    - BE 12 % CZ 6 % DE 12 % GR 25 % SK 12 %
  - Standard drop-outs
    - BE 56 % CZ 44 % DE 48 % GR 68 % SK 58 %
- Not interested
  - Greeks stand out, unfortunately the sample set too small
  - There is a remarkable number of really high bids
    - creating “linearity” from “not interested” to average bid

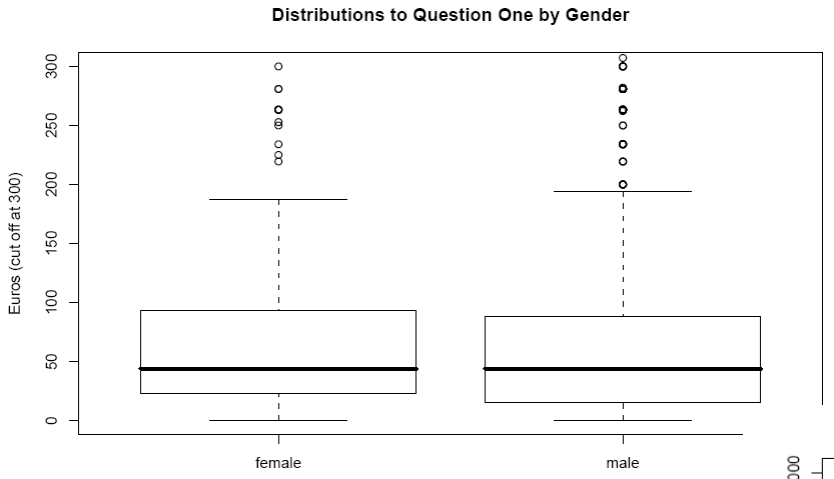
Differences among Countries

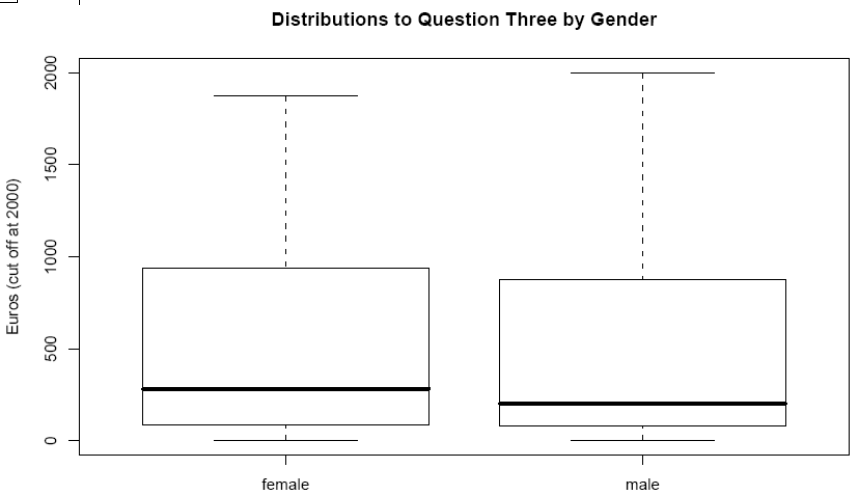
- 1st bids





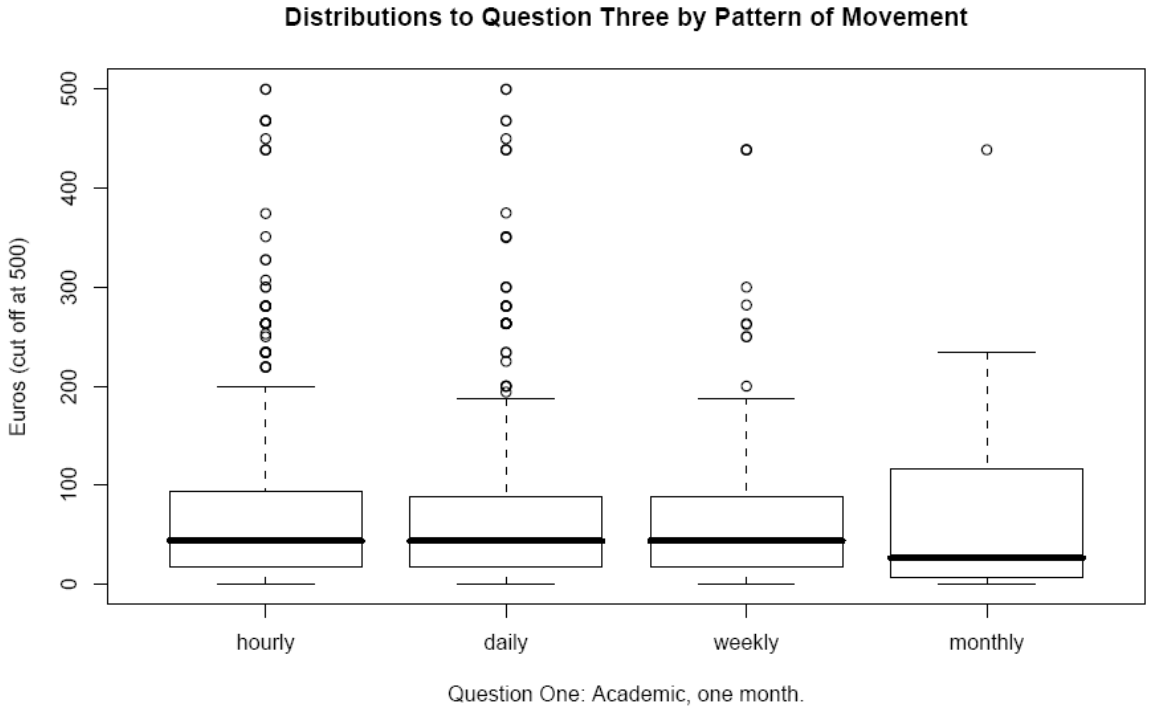
Men and Women





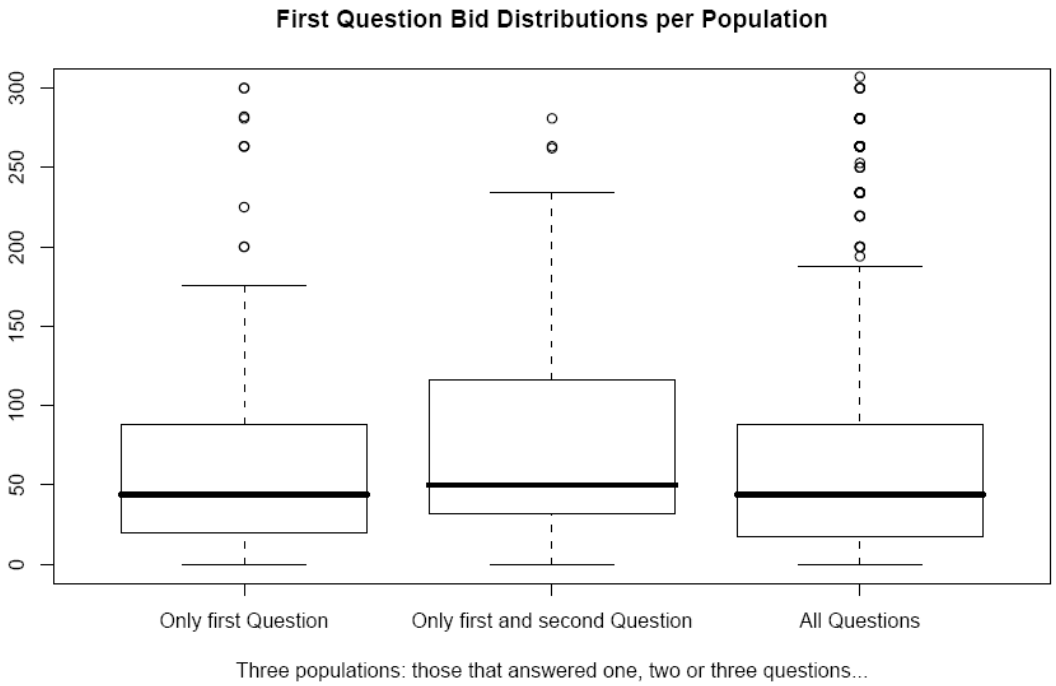
- Medians of the 2nd bids
  - 1.4 : 1
- Medians of the 3rd bids
  - 1.8:1

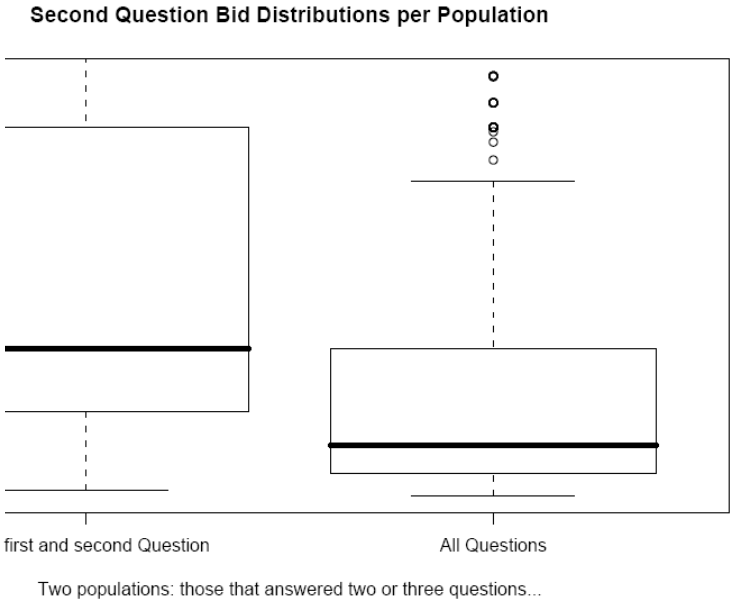
Mobility



- Sizes of sample sets: daily 520, hourly 485, weekly 195, monthly 15
- Expectation was that there is correlation between value of irregular movements

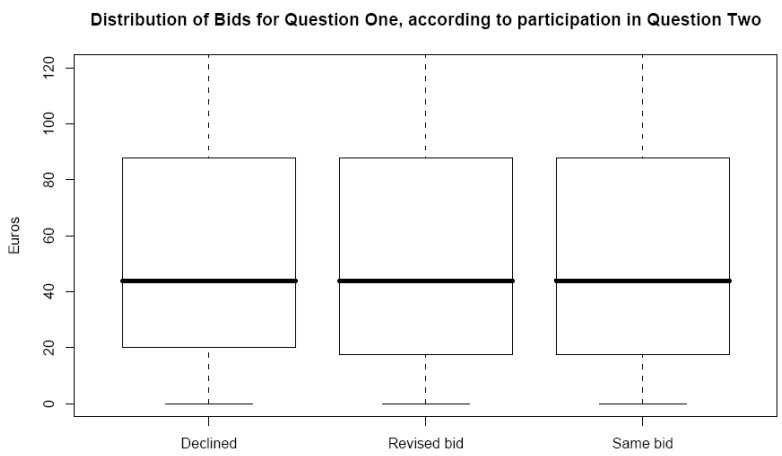
Impact of Scenarios





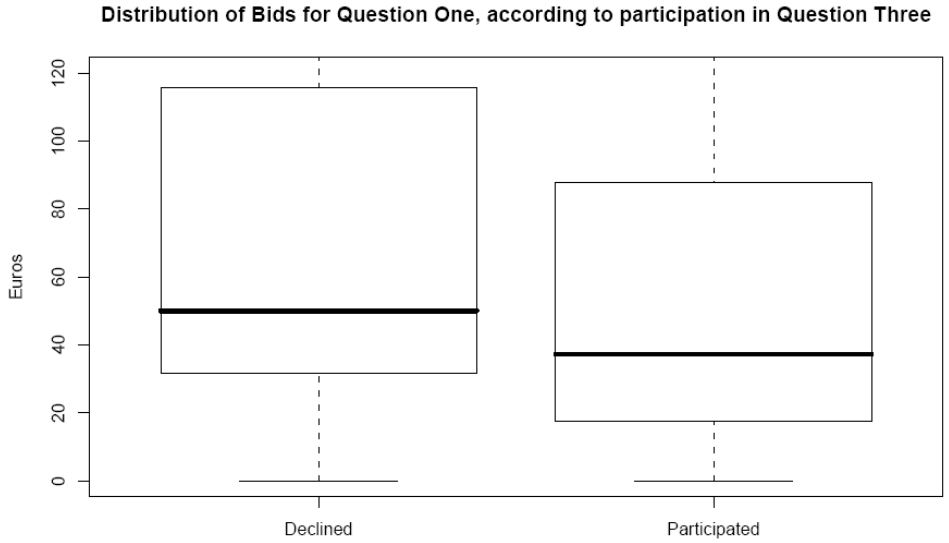
- Curiosity vs privacy cautiousness
  - Left – low bids: curiosity and falling off in the second round
  - Middle – higher bids, increased in the second round
  - Right – low first bid increased in each consecutive round

Impact of Scenarios II

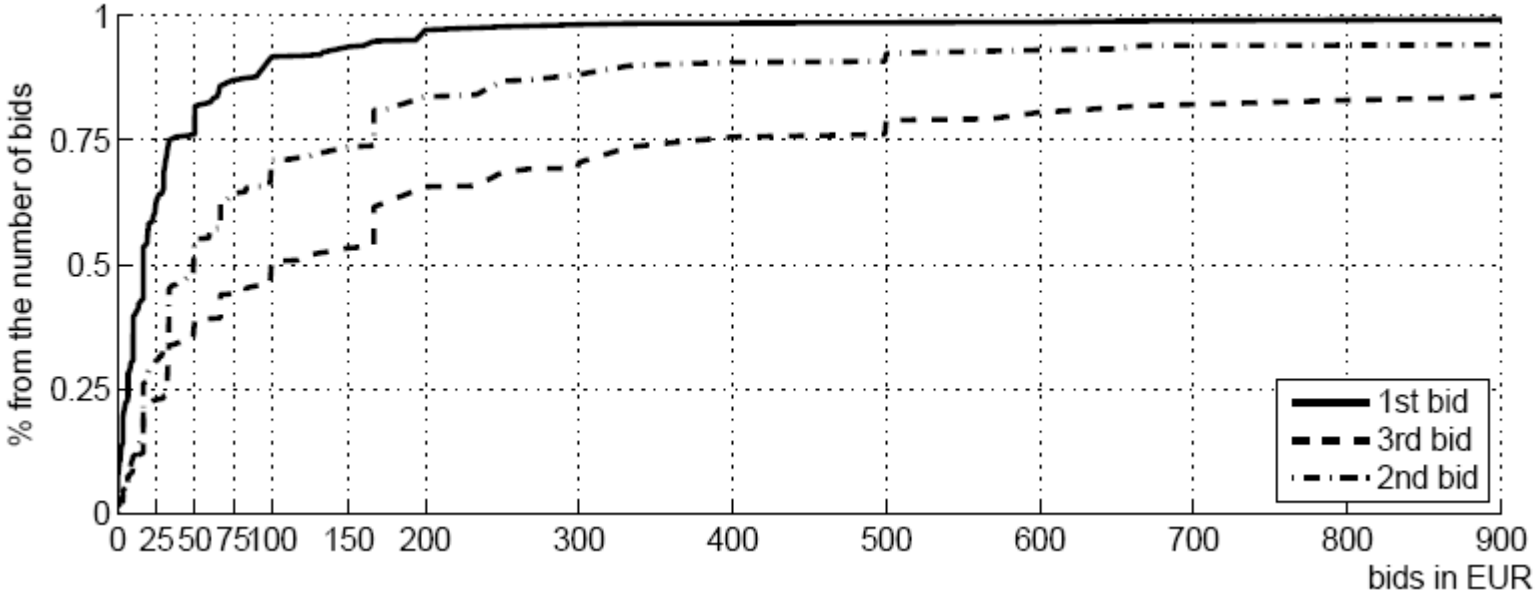


Value of bids according to answers in the second round (decline, same, revise)

Bids according to answers
In the third round (declined, participated)



Overall Distribution of Bids

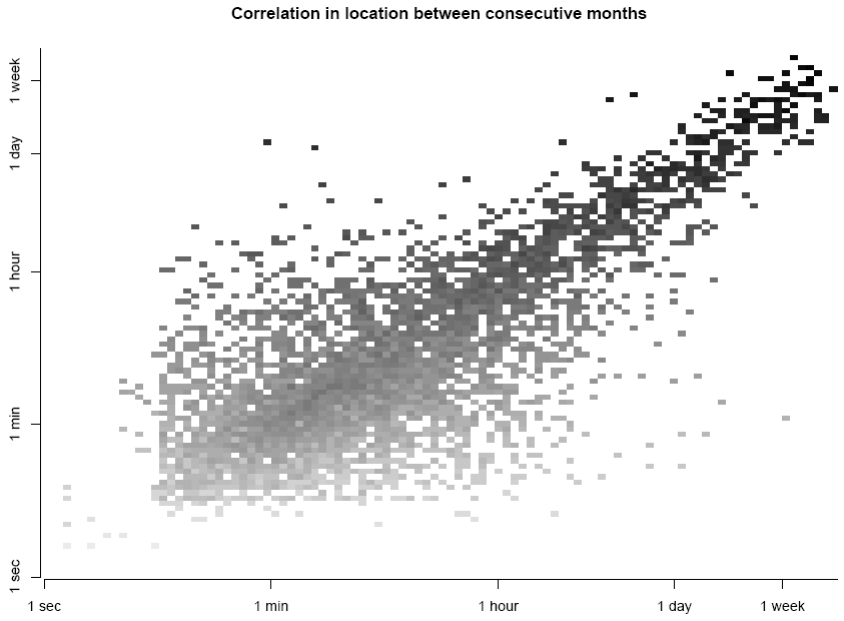


- Second bids (2.5x first bids)
- Third bids (2x second bids)

Non-linearity in Time

- 12-fold increase in the experiment length
  - 2x increase of the bids
- Hypothesis
  - Data after the 1st month are of less value
  - Little information in consecutive data

Correlation between consecutive months (MIT Reality Mining project)



Why Participating in The Study

- Questioned after the experiment
  - 300 responses (25 % of the participants)
- Why did you take part in the experiment
  - Money (38 %), results (32 %), fun (30 %)
- Correlated bid values (medians)
  - 1^{st} auction: 12, 8, 9 (roughly)
  - 2^{nd} auction: 9, 5, 6 (roughly)
  - No substantial difference between bids

Conclusions

- 10 % of participants bidding < 1 EUR
  - Curiosity and enthusiasm for cover story
- Greek sensitivity to privacy breaches
  - Eavesdropping scandal a couple of months before
- Non-linearity in regard of the study length
- No correlation between bids and movements
- Medians of Cambridge study correspond to our results (€43 to £28)

Introduction – second study

- Usage of online communication tools
- Email or instant messaging used every day
- Network administrators can track their users
- Risk of profiling or another analyzes of data
- People can sense the value of such information

Organisation of the study

- How much money for being tracked for two weeks
  - email
  - instant messaging
  - all tracking data
- **First form** (webpage) – do you want to take part?
  - **Academic research**
  - Yes, with a PC only
- Second form – partially supporting our cover story
  - Age?, Gender?
  - Own or shared hardware?
  - Level of IT-knowledge?
- **Second bid – commercial exploitation** (decline, revised bid)
- **Third bid – use by national governments** to improve terrorist activity detection and tracking tools

Structure of responders

- Intent to participate in the first step (academic research usage of data) of the study – 498 subjects (of 1080 loads)
  - BE(3.4%), CZ(40.2%), DE(8%),
  - SK(32.1%), EN(16.3%)
- 284 then actually bid (first scenario)
- Those who saw the introtext and answered
  - will participate – 46.1 %, (26.3 % – first scenario)

Academic usage (quartiles)

First bids			First bids – males			First bids - females		
email	messaging	all	email	messaging	all	email	messaging	all
10	10	12	10	9.5	12	10	10	15
30	30	50	32.5	25	50	30	35	50
100	100	200	100	100	200	275	150	300

- Quartiles instead of min, max, average values
- 23 participants (almost 10%) explicitly opt out for the next scenario, but 27% left

Commercial usage

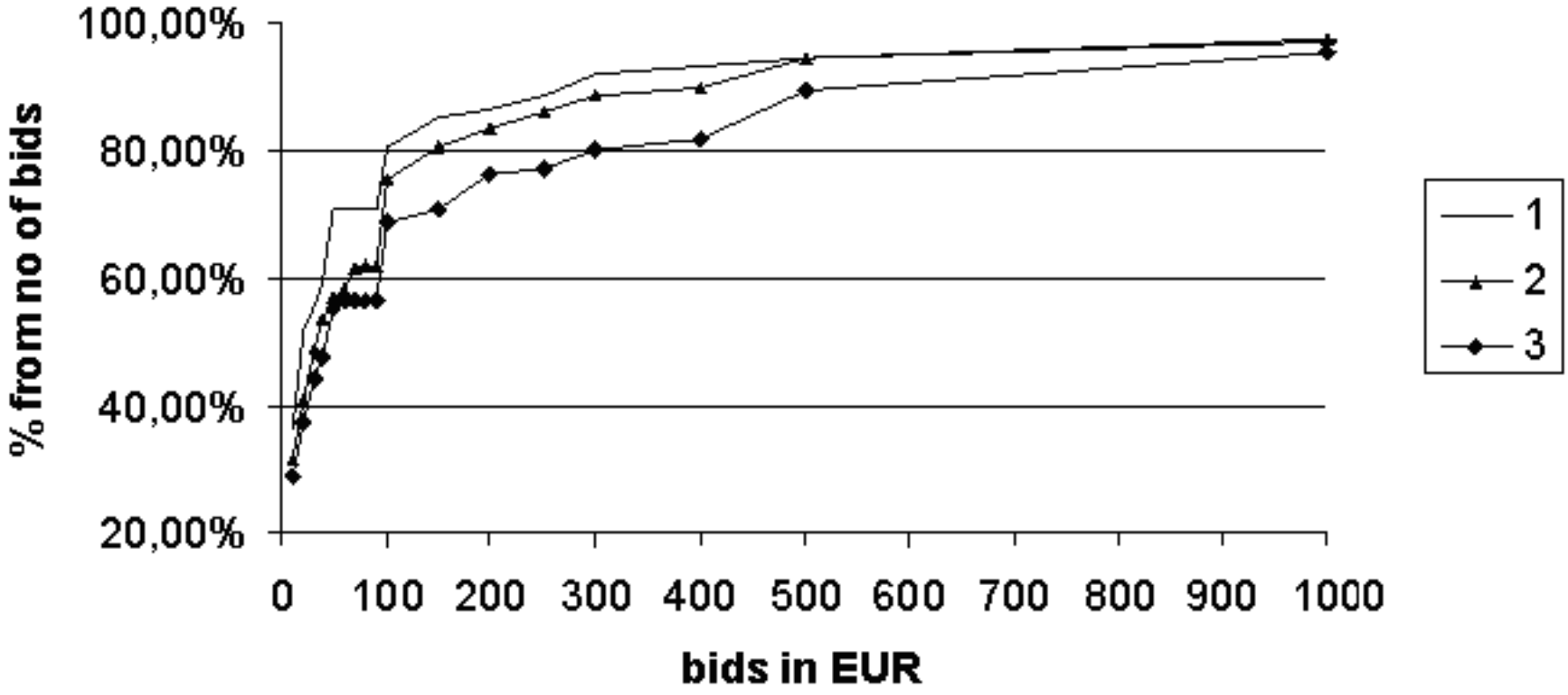
Academic			Commercial			% increase
email	messaging	all	email	messaging	all	
10	8.3	10.4	10	10	15	22%
20	22.5	40	40	40	50	57%
100	80	150	100	100	200	21%

- Medians increased significantly
- 41 participants (18%) explicitly opt out in the next scenario, but 28% actually left

Usage by governments

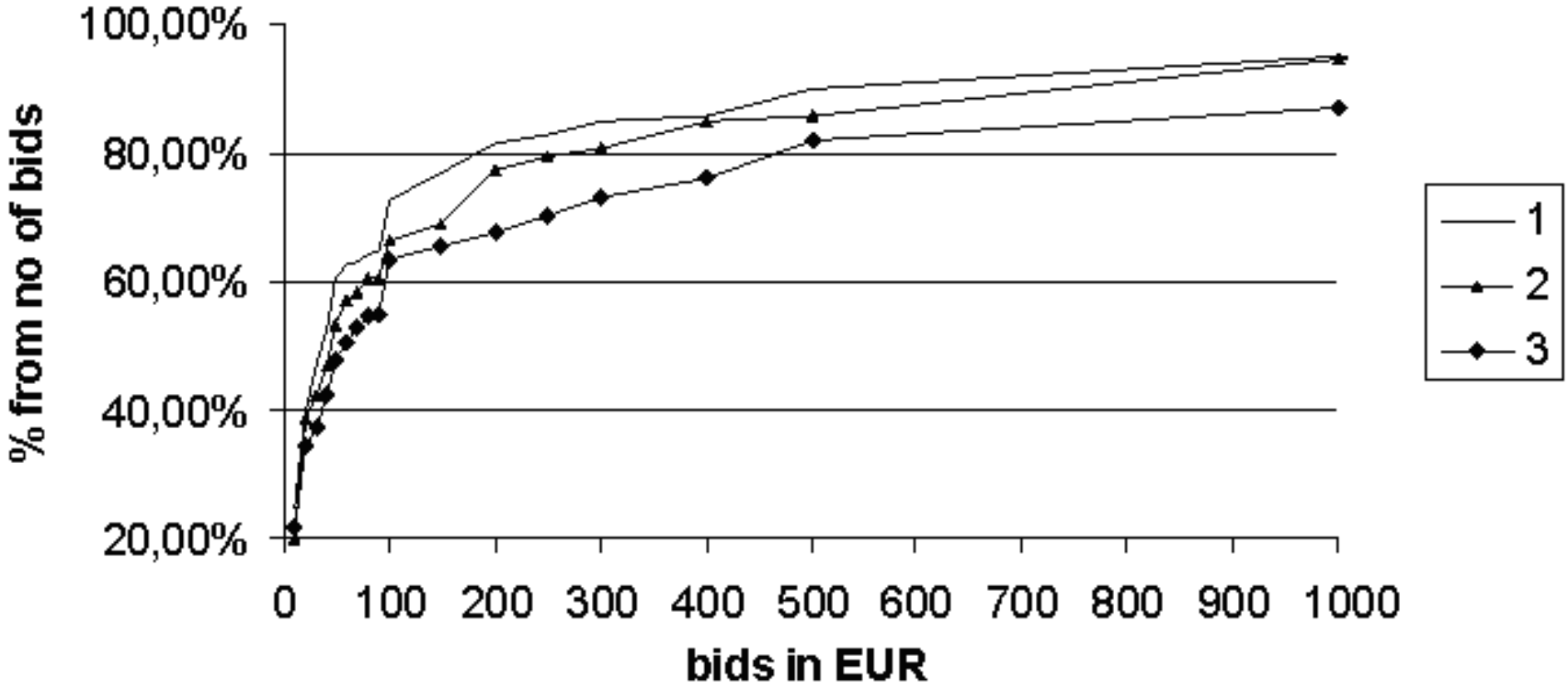
Second bids			Third bids		
email	messaging	all	email	messaging	all
10	10	15	10	10	15
40	40	50	50	50	60
100	100	200	200	200	400

Histogram – 1st bid, all scenarios



- Higher differences expected

Histogram – 3rd bid, all scenarios



Highlights of the second study

- 284 responses for at least the first scenario
  - responses from more than four countries
- €30 for being tracked (email or instant messaging) for academic purposes
  - €50 for all tracking data
  - No considerable differences between males and females
- Increasing tendency to opt out with changing purpose of tracking
  - 1/10 academic -> commercial usage (real dropout 27%)
  - 1/5 commercial -> governmental usage (real dropout 28%)
- Governmental usage
  - After dropouts, ie valuation of all-consenting subjects
  - €50 for one type of data (cf. €40 commercial, €20/25 acad.)
- No significant difference between value of email and other messaging traffic data

Rozsáhlé databáze osobních informací

Vašek Matyáš

PV080

Agregace dat

- Seskupování (osobních) dat do rozsáhlých databází. Agregace (z angl. *aggregation*).
- Tímto kombinováním dat o určité citlivosti lze získat informace daleko citlivější, které jinak spadají do kategorie s vyššími požadavky na ochranu.

Zákon o ochraně osobních údajů (101/2000 Sb.) – Povinnosti správce

Mj. zákon říká:

- nesdružovat osobní údaje, které byly získány k rozdílným účelům, pokud zvláštní zákon nestanoví jinak

Žadatel o investici

- Chodil roky ke stejnému obvodnímu lékaři.
- Uzavřel před měsícem vysokou živ. pojistku.
- V minulém čtvrtletí byl u specialisty.
- Přede dvěma měsíci změnil obvodního lékaře.

Odvození (Inference, i angl.)

- Odvození informací o vyšší citlivosti zpracováním a analýzou skupiny informací o nižší citlivosti.

nebo

- Nepřímý přístup k informacím bez přímého přístupu k datům, která tyto informace reprezentují.

Příklad politiky klinických IS, British Medical Association

- Musí být zavedena účinná opatření proti agregaci osobních zdravotních informací.
- Pacienti, k jejichž seznamu řízení přístupu má být přidána další osoba, musí být zvlášť upozorněni, pokud již tato osoba má přístup ke zdravotním informacím velkého množství lidí.

Co když máte informace o finanční situaci a zdrav. stavu

1. Přítele/kyně, resp. manžela/ky.
2. Spolupracovníka, nadřízeného...
3. Všech studentů/zaměstnanců FI.
4. Všech obyvatel místa, kde žijete.
5. Všech klientů určité firmy (banky, zdravotní pojišťovny...).
6. Všech (většiny) občanů.

Pravděpodobnost neoprávněného použití

- Počet osob, které mají k informacím přístup (operátoři, uživatelé systému ap.).
- Hodnota informací.
  - Výše trestu těm, kdo data jiných řádně neohlídali a spolupodíleli se tak na jejich úniku.
  - Výše trestu těm, kdo s nimi neoprávněně manipulují.
  - Úroveň ochranných mechanismů.

Řešení?

- U menších souborů osobních dat provádět agregaci jen v nutných případech.
- U větších souborů neprovádět agregaci.

- Statistické databáze!

Statistické databáze

- Obsahují citlivé údaje o jednotlivcích.
- Jejich využití má být **jen** pro statistické dotazy k vytvoření obrazu o celkových potřebách obyvatelstva a formulování (vládní) politiky.
  - podpora církví, regionů/měst atd.
- Výsledky dotazů v takovýchto databázích nesmějí poskytnout údaje o jednotlivcích.

Studium statistických databází

- USA, 70. léta, databáze ze sčítání lidu.
- Dorothy Denning
  - Studium používaných způsobů pro formulaci dotazů a získávání odpovědí.
  - Ty povolovaly (netriviální!) dotazy, které umožnily získat údajně tajné informace o jednotlivci.
  - ☺ Údajně nedůvěra ve zjistění Denningové – dokud nezjistila plat svého šéfa sérií legitimních dotazů.

Příklad kritického dotazu

Kolik je měst s 15-16 000 obyvatel

& s muži, evangelíky, slovenské nár., 36-40 let

& jejich ženy, 28-30 let žijí mimo toto město

& 2 děti do 10 let žijí s těmito ženami

& 1 dítě nad 18 žije s těmito muži

& muž žije ve vlastním domě, plocha nad 200m^2
a domácnost má/používá aspoň 2 automobily.

Kompromitace databáze

- Výsledkem série dotazů je jeden záznam
  - Databáze byla pozitivně kompromitována
- Následný pokus o získání dalších informací
  - Výsledkem je buď 1 nebo 0 záznamů
  - Pozitivní/částečná kompromitace databáze
- Částečná kompromitace
  - Informace o entitě i když neznáme konkrétní hodnotu

Protiopatření ve statistických databázích I.

- Omezení dotazu
  - Např. i sledování předchozích dotazů
- Úmyslná změna zdrojových dat
  - Např. orig. hodnoty nahrazeny novým vzorkem se stejným rozložením pravděpodobnosti hodnot
- Úmyslná změna výsledku dotazu
  - Např. zaokrouhlování
- Cílem je zabránit situacím, kdy je možné získat informace o jedné entitě.

Protiopatření ve statistických databázích II. – *Náhodný výběr*

Každý dotaz je zodpovězen na základě vyhodnocení náhodně vybraných záznamů ze všech existujících záznamů.

- Kontrola překrytí množiny záznamů u vícenásobných dotazů na tutéž informaci.
  - Má zabránit situaci, kdy několik uživatelů databáze začne spolupracovat.
- Technika nyní používaná v americké databázi údajů ze sčítání lidu.

Protiopatření ve statistických databázích III. – *Minimální rozsah dotazu*

- Minimum celkového počtu záznamů použitých pro tvorbu odpovědí.

nebo

- Minimum počtu záznamů použitých pro tvorbu odpovědí na každou část dotazu.

Protiopatření ve statistických databázích III. – *Perturbační (zmatečné) techniky*

Přidání pseudonáhodného „šumu“:

- Odpovědi konzistentní, ale získání spolehlivé odpovědi na sérii podobných dotazů není možné.

1. K záznamům zahrnutým pro vyhodnocení dotazů se přidají další náhodně vybrané podobné záznamy
2. Vypočtená hodnota nebo mezihodnoty jsou zaokrouhlovány nebo mírně pozměněny.

- Podle některých definic zahrnují *náhodný výběr*.

De-anonymizace uživatelů

- Narayanan a Shmatikov (2008)
  - Huge de-anonymization of large sparse datasets (ACM)
- Databáze hodnocení filmů
  - Databáze zpřístupněna „anonymizovaně“
- Uživatel hodnotí filmy (filmů jsou stovky) na škále 1-10
- Uživatele se podařilo de-anonymizovat – spojit se skutečnou identitou pokud:
  - Víme jeho hodnocení pro 5-8 filmů