CheckUser and Editing Patterns

Balancing privacy and accountability on Wikimedia projects

Wikimania 2008, Alexandria, July 18, 2008

HaeB
haebwiki@gmail.com
[[de:Benutzer:HaeB]], [[en:User:HaeB]]

Please don't take photos during this talk.

What are sockpuppets?

- Wikipedia and sister projects rely on being open: Anyone can edit, anyone can get an account.
- No ID control when registering (not even email address required)
- Many legitimate uses for multiple accounts
- “Sockpuppet” often implies deceptive intention, think ventriloquist

What is the problem with sockpuppets?

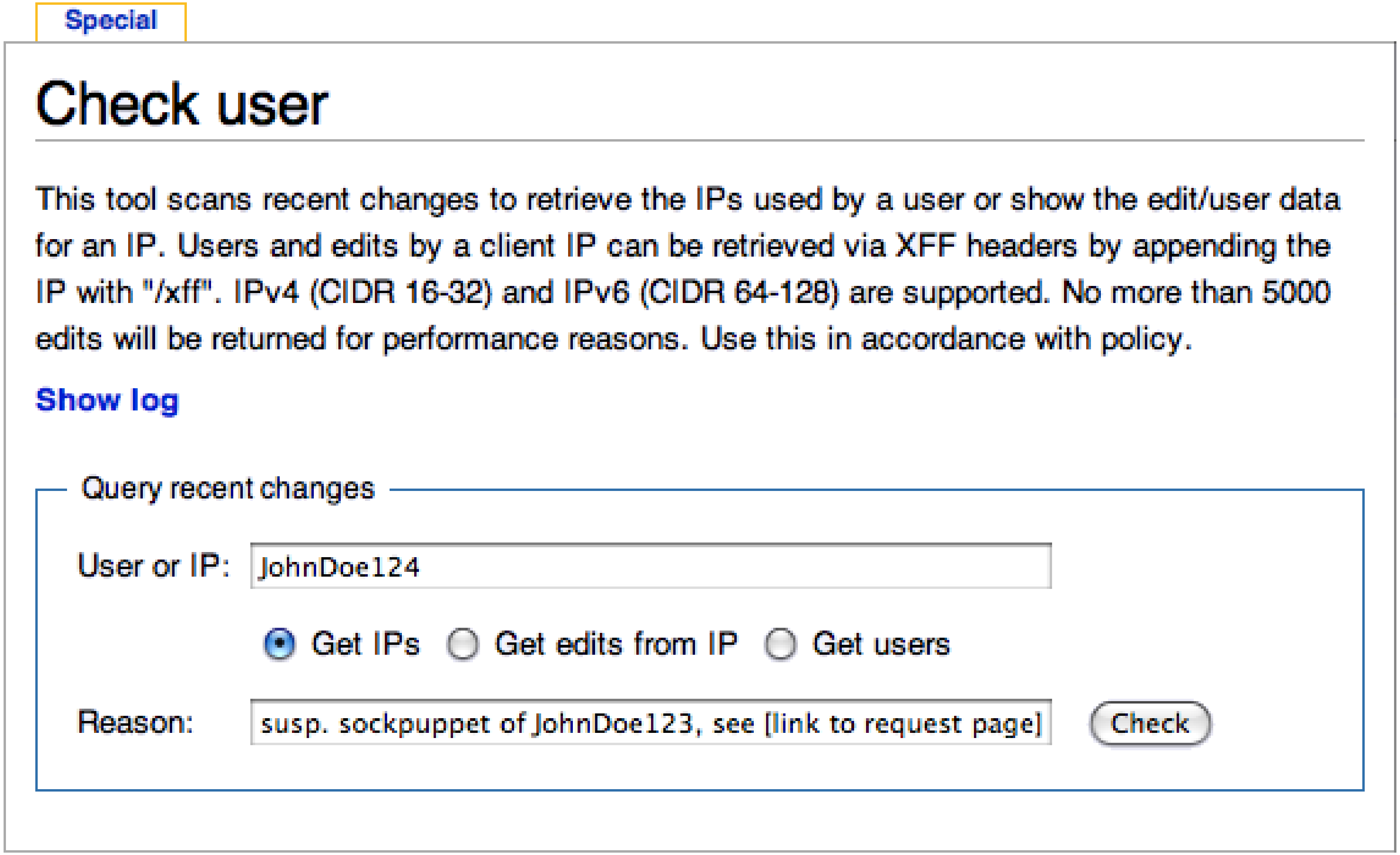
- Ballot stuffing (some decision processes rely on voting, such as request for adminships and WMF board elections)
- “Dr Jekyll/Mr Hyde”: Carry out evil or controversial actions with a sockpuppet, such that the main account remains in good standing. E.g. trolling (actions intended to provoke adversive reactions and disrupt the community), or strawman accounts (putting the adversarial position in bad light)
- Artificial majorities in content disputes (useful if the wiki's culture values majority consensus over references and arguments), especially circumventing “three-revert-rule”
- Ban evasion

What is the problem with sockpuppets?

(cont'd)

- Newbies get bitten as a result of the possibility of ban evasion:
    - Friedman and Resnick (*The Social Cost of Cheap Pseudonyms*, Journal of Economics and Management Strategy 2001): Proved in a game-theoretic model that the possibility of creating sockpuppets leads to distrust and discrimination against newcomers (if the community is otherwise successfully cooperating as a whole)
- Summarily: The reputation system of an online community relies on accounting actions, sockpuppets disrupt this.

The CheckUser tool in MediaWiki



Check user

This tool scans recent changes to retrieve the IPs used by a user or show the edit/user data for an IP. Users and edits by a client IP can be retrieved via XFF headers by appending the IP with "/xff". IPv4 (CIDR 16-32) and IPv6 (CIDR 64-128) are supported. No more than 5000 edits will be returned for performance reasons. Use this in accordance with policy.

Show log

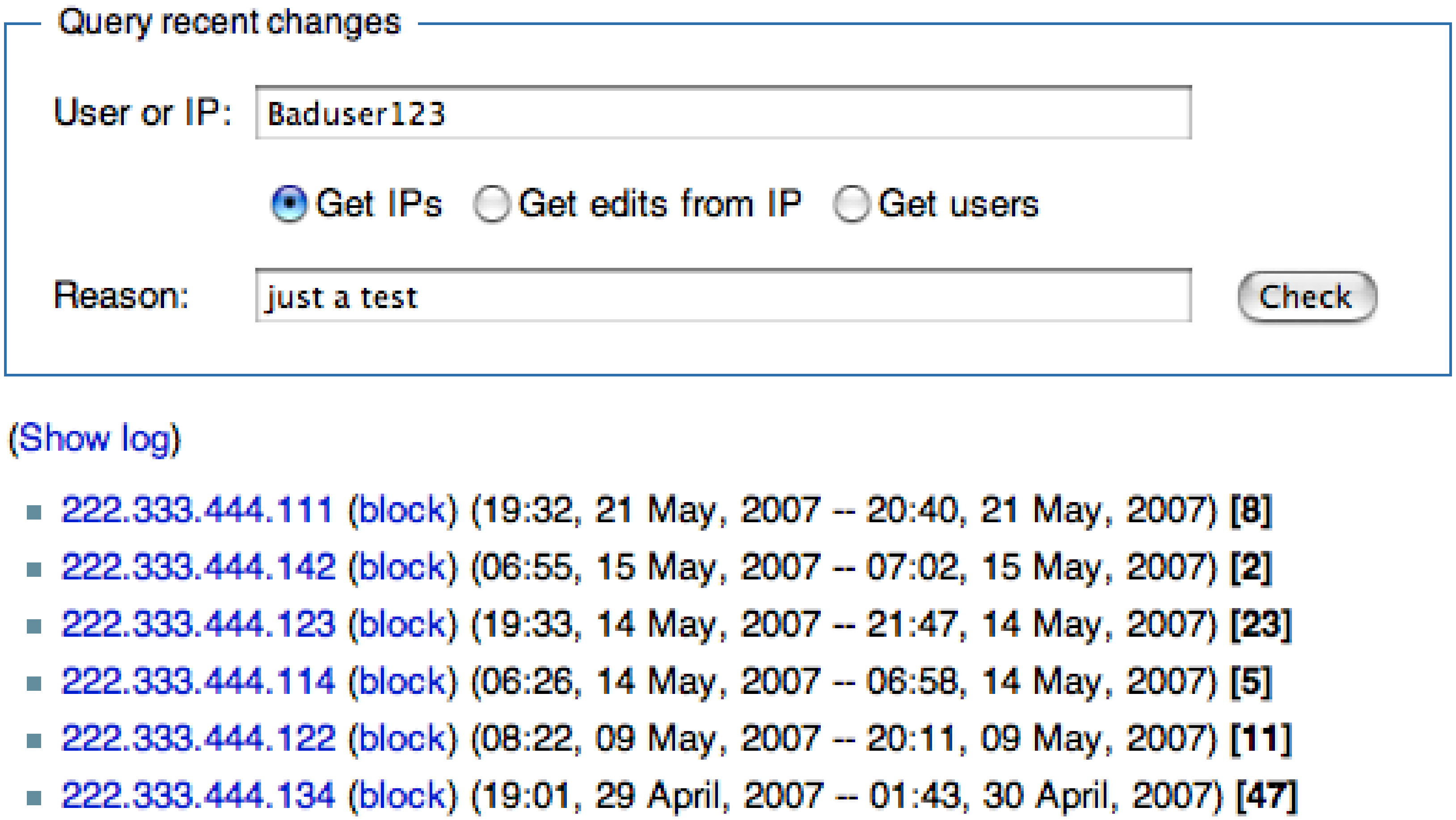
Query recent changes

User or IP: JohnDoe124

● Get IPs ○ Get edits from IP ○ Get users

Reason: susp. sockpuppet of JohnDoe123, see [link to request page]

Get IPs of a logged-in user



Query recent changes

User or IP: Baduser123

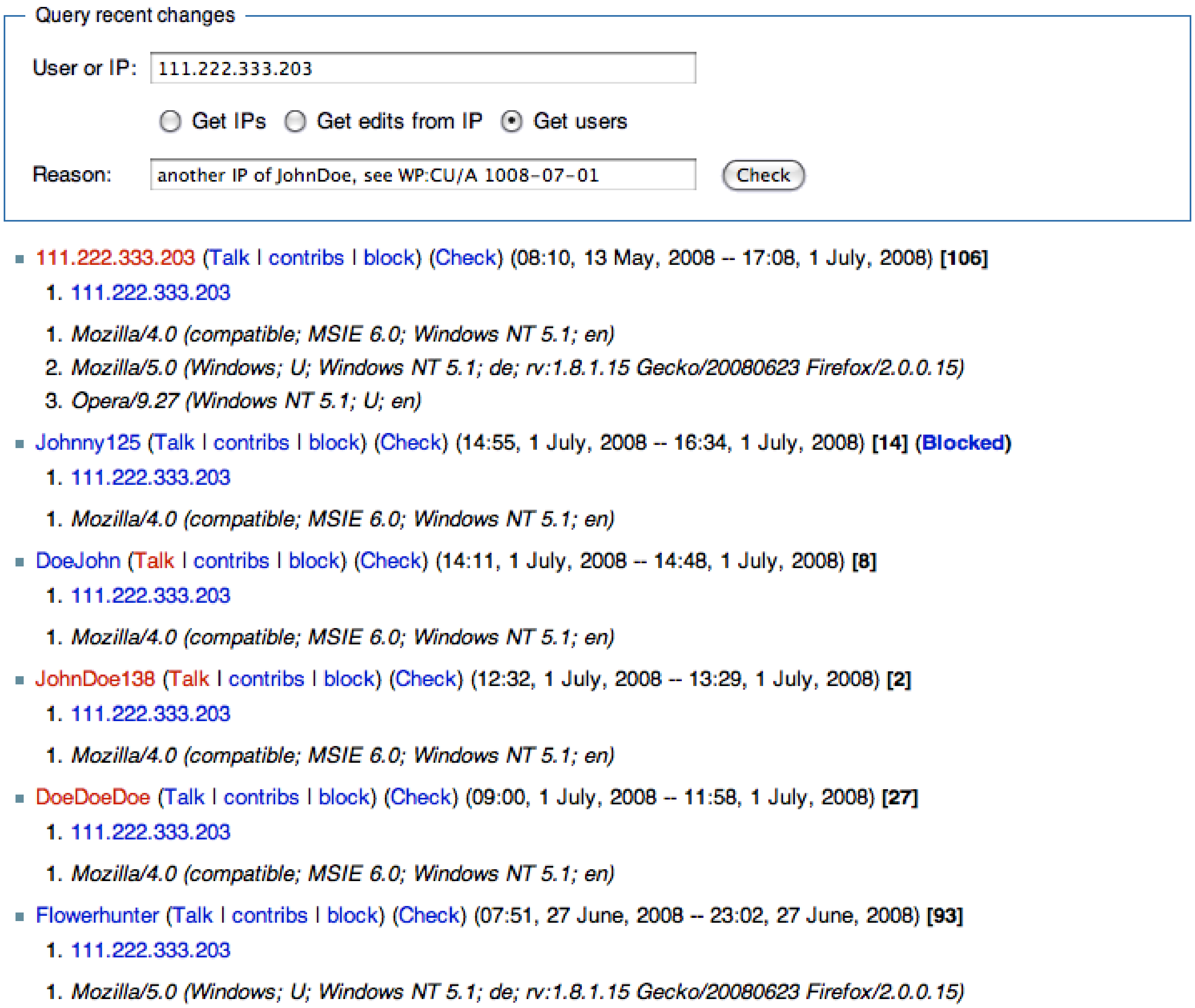
● Get IPs ○ Get edits from IP ○ Get users

Reason: just a test

(Show log)

- 222.333.444.111 (block) (19:32, 21 May, 2007 -- 20:40, 21 May, 2007) **[8]**
- 222.333.444.142 (block) (06:55, 15 May, 2007 -- 07:02, 15 May, 2007) **[2]**
- 222.333.444.123 (block) (19:33, 14 May, 2007 -- 21:47, 14 May, 2007) **[23]**
- 222.333.444.114 (block) (06:26, 14 May, 2007 -- 06:58, 14 May, 2007) **[5]**
- 222.333.444.122 (block) (08:22, 09 May, 2007 -- 20:11, 09 May, 2007) **[11]**
- 222.333.444.134 (block) (19:01, 29 April, 2007 -- 01:43, 30 April, 2007) **[47]**

Get edits



Query recent changes

User or IP: 111.222.333.203

○ Get IPs ○ Get edits from IP ● Get users

Reason: another IP of JohnDoe, see WP:CU/A 1008-07-01

- 111.222.333.203 (Talk | contribs | block) (Check) (08:10, 13 May, 2008 -- 17:08, 1 July, 2008) **[106]**
  1. 111.222.333.203

  1. *Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; en)*
  2. *Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.15 Gecko/20080623 Firefox/2.0.0.15)*
  3. *Opera/9.27 (Windows NT 5.1; U; en)*
- Johnny125 (Talk | contribs | block) (Check) (14:55, 1 July, 2008 -- 16:34, 1 July, 2008) **[14]** **(Blocked)**
  1. 111.222.333.203

  1. *Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; en)*
- DoeJohn (Talk | contribs | block) (Check) (14:11, 1 July, 2008 -- 14:48, 1 July, 2008) **[8]**
  1. 111.222.333.203

  1. *Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; en)*
- JohnDoe138 (Talk | contribs | block) (Check) (12:32, 1 July, 2008 -- 13:29, 1 July, 2008) **[2]**
  1. 111.222.333.203

  1. *Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; en)*
- DoeDoeDoe (Talk | contribs | block) (Check) (09:00, 1 July, 2008 -- 11:58, 1 July, 2008) **[27]**
  1. 111.222.333.203

  1. *Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; en)*
- Flowerhunter (Talk | contribs | block) (Check) (07:51, 27 June, 2008 -- 23:02, 27 June, 2008) **[93]**
  1. 111.222.333.203

  1. *Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.15 Gecko/20080623 Firefox/2.0.0.15)*

- Information available for each edit:
  - IP address under which the edit was made
  - User agent (browser version, operating system version)
  - XFF (X-Forwarded-For) data: If the editor used a proxy which supports it (most don't), shows originating IP too
- Information ***not*** shown: Email address or other account settings, user's password, screen resolution, browser plugins ...
- CheckUser information only stored for a limited time (currently 90 days), checks for older edits not possible
- Besides sockpuppet investigations, other applications (e.g. finding the IP range used by for heavy, repeated vandalism, to enable a range block)

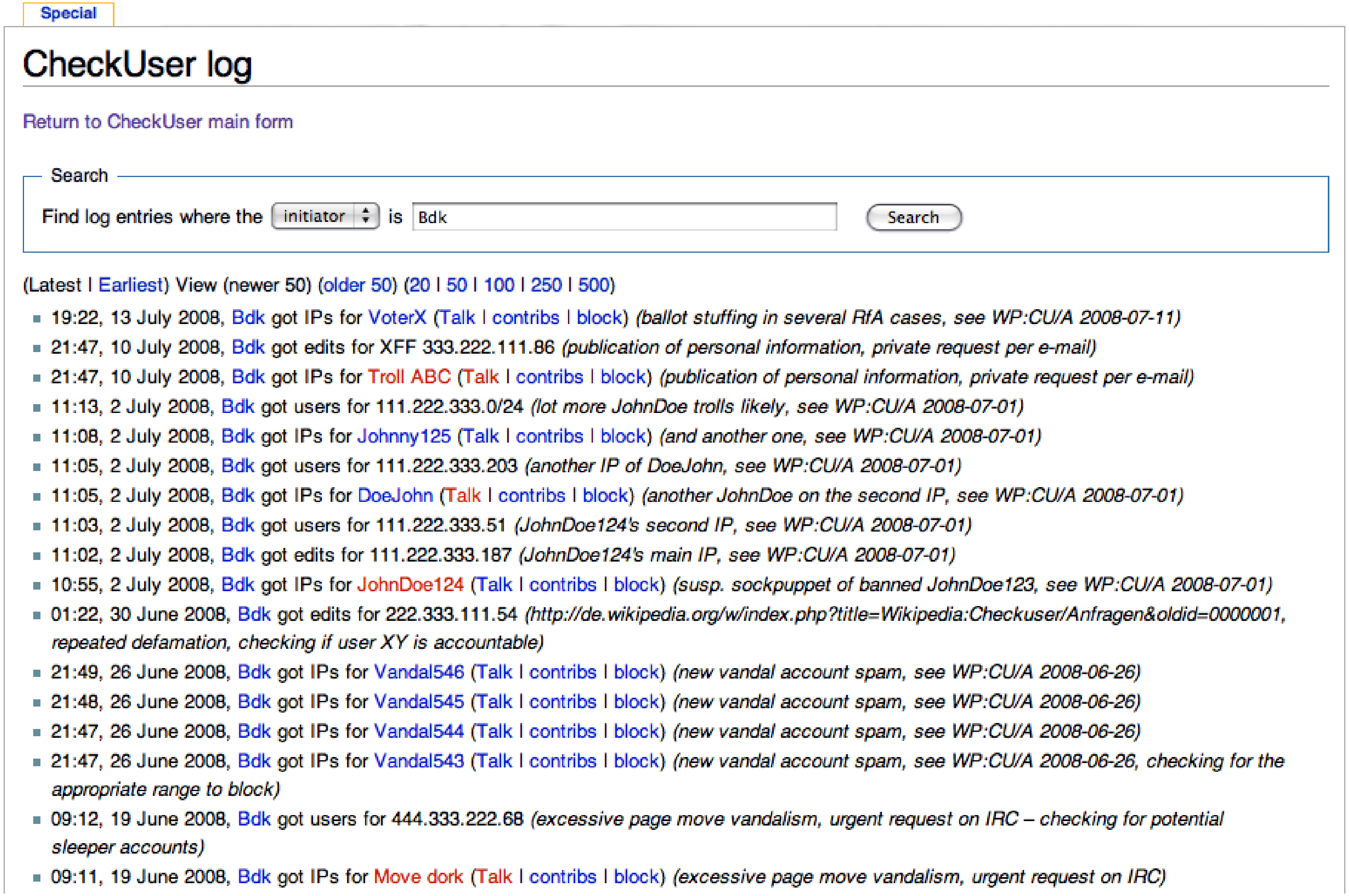
Some history

- During the first years, developers (server admins) did checks by hand on request
- CheckUser introduced as a MediaWiki extension by Tim Starling in early 2005
- Trivia: A predecessor was called “Espionage“
- Fall 2005: WMF CheckUser policy
- 2006: WMF introduced CheckUser ombudsperson commission
- 2007: Added user agents and XFF
- 2008: New privacy policy

Who gets access to CheckUser data?

- For privacy reasons, only a few trusted users
- CheckUser rights granted by community vote or assigned by ArbCom - process varies between projects
- Must be over 18 years and identify to WMF
- Stewards can give themselves CheckUser rights to carry out cross-wiki checks
- CheckUser ombudsmen can view all logs
- Developers still have access to the same information, too
- Coordination (e.g. cross-wiki checks) via a closed mailing list and a chatroom

CheckUser log



CheckUser log

Return to CheckUser main form

Search

Find log entries where the initiator is Bdk

(Latest | Earliest) View (newer 50) (older 50) (20 | 50 | 100 | 250 | 500)

- 19:22, 13 July 2008, Bdk got IPs for VoterX (Talk | contribs | block) *(ballot stuffing in several RfA cases, see WP:CU/A 2008-07-11)*
- 21:47, 10 July 2008, Bdk got edits for XFF 333.222.111.86 *(publication of personal information, private request per e-mail)*
- 21:47, 10 July 2008, Bdk got IPs for Troll ABC (Talk | contribs | block) *(publication of personal information, private request per e-mail)*
- 11:13, 2 July 2008, Bdk got users for 111.222.333.0/24 *(lot more JohnDoe trolls likely, see WP:CU/A 2008-07-01)*
- 11:08, 2 July 2008, Bdk got IPs for Johnny125 (Talk | contribs | block) *(and another one, see WP:CU/A 2008-07-01)*
- 11:05, 2 July 2008, Bdk got users for 111.222.333.203 *(another IP of DoeJohn, see WP:CU/A 2008-07-01)*
- 11:05, 2 July 2008, Bdk got IPs for DoeJohn (Talk | contribs | block) *(another JohnDoe on the second IP, see WP:CU/A 2008-07-01)*
- 11:03, 2 July 2008, Bdk got users for 111.222.333.51 *(JohnDoe124's second IP, see WP:CU/A 2008-07-01)*
- 11:02, 2 July 2008, Bdk got edits for 111.222.333.187 *(JohnDoe124's main IP, see WP:CU/A 2008-07-01)*
- 10:55, 2 July 2008, Bdk got IPs for JohnDoe124 (Talk | contribs | block) *(susp. sockpuppet of banned JohnDoe123, see WP:CU/A 2008-07-01)*
- 01:22, 30 June 2008, Bdk got edits for 222.333.111.54 *(http://de.wikipedia.org/w/index.php?title=Wikipedia:Checkuser/Anfragen&oldid=0000001, repeated defamation, checking if user XY is accountable)*
- 21:49, 26 June 2008, Bdk got IPs for Vandal546 (Talk | contribs | block) *(new vandal account spam, see WP:CU/A 2008-06-26)*
- 21:48, 26 June 2008, Bdk got IPs for Vandal545 (Talk | contribs | block) *(new vandal account spam, see WP:CU/A 2008-06-26)*
- 21:47, 26 June 2008, Bdk got IPs for Vandal544 (Talk | contribs | block) *(new vandal account spam, see WP:CU/A 2008-06-26)*
- 21:47, 26 June 2008, Bdk got IPs for Vandal543 (Talk | contribs | block) *(new vandal account spam, see WP:CU/A 2008-06-26, checking for the appropriate range to block)*
- 09:12, 19 June 2008, Bdk got users for 444.333.222.68 *(excessive page move vandalism, urgent request on IRC – checking for potential sleeper accounts)*
- 09:11, 19 June 2008, Bdk got IPs for Move dork (Talk | contribs | block) *(excessive page move vandalism, urgent request on IRC)*

CheckUser log

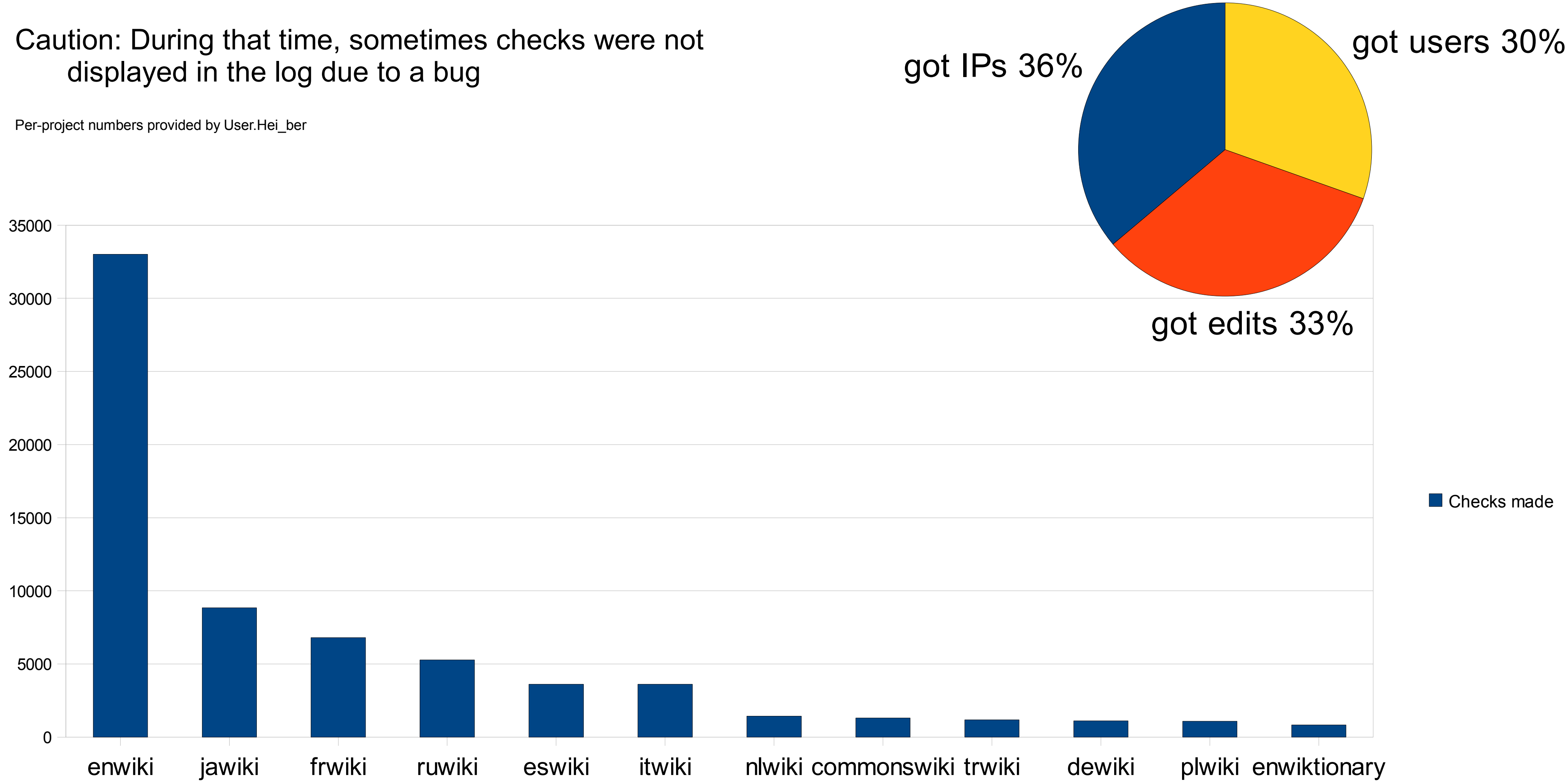
- Visible to other CheckUsers on the same projects, CheckUser ombudsmen and developers
- Used to be visible to CheckUsers from all projects (global log, has been disabled for some time due to technical problems)

Some usage statistics

Source: Global CheckUser log, October 2006-December 2007

Caution: During that time, sometimes checks were not displayed in the log due to a bug

Per-project numbers provided by User.Hei_ber



Interpreting CheckUser results

- Naively:
  - “Account A uses the same IP as editor B, therefore A and B are the same person.”
  - “A and B use different IPs, therefore A and B are different persons.”
- Wrong for several reasons:
  - People don't always use the Internet from the same entry point (travel, home/work, ...)
  - The same router (entry point) is often used by more than one computer (LANs of companies, families, flatmates...)
  - The same router is often assigned different IPs over time: it has a “dynamic IP” (especially common with DSL and phone dialup)

Interpreting CheckUser results – a formal approach

- If account B (suspected sockpuppet) uses the same IP or same dynamic IP range as account A, how sure is it that they are the same person (A = B)?
- General question, written using ***conditional probabilities***:
  - P(H|E) = ? where
  - H is the hypothesis (“A=B”)
  - E is the evidence (both use the same IP range to access the Internet)
  - P(H|E) is the probability for H occurring if we know that E has occurred (“conditional probability”)
  - Assuming B is from the group of all Internet users, with no further knowledge

Bayes' Theorem

$$P(H|E) = P(E|H)\,\frac{P(H)}{P(E)}$$

- Simple example:
  - B generated by throwing a fair die (B = ⚀,⚁,⚂,⚃,⚄,⚅)
  - Hypothesis H: “B is a ⚄” (i.e. A=⚄ in above notation)
  - Evidence E: “B is odd”
  - P(E|H) = 1 (because 5 is always an odd number), P(E) = ½ (half of all numbers are odd), P(H) = 1/6 (die is fair)
  - With Bayes: P(H|E) = 1/3
- Very frequently applied in forensic statistics (DNA evidence, etc.)

$$P(H|E) = P(E|H)\frac{P(H)}{P(E)}$$

Bayes' Theorem applied to CheckUser

- Reminder: B is assumed to come randomly from the crowd of all Internet users (“just some random surfer”). Say that there are 1 billion of them, then $P(H)=1/1 \text{ billion}=10^{-9}$
- Say the shared IP range is XXX.YYY.0.0-XXX.YYY.63.255 (i.e. 2^{14}=16384 different IPs). Assume for simplicity that IP addresses are evenly distributed from 0.0.0.0 to 255.255.255.255 (not actually true). Then $P(E)=2^{14}/2^{32}=2^{-18}$
- As in the die example, P(E|H)=1 (i.e. “no false negatives”: we chose the IP range to encompass all of the IPs that A uses)
- From Bayes, we get $P(H|E) = P(H)/P(E) \approx 10^{-9}2^{18} \approx 0.026\%$. Not very impressive :(
- But we haven't used all our knowledge: We know that A and B have both edited this wiki (not all Internet users do!)

$$P(H|E) = P(E|H)\,\frac{P(H)}{P(E)}$$

Combining evidence

- Instead of E = “A and B use the same IP range”, consider $E = (E_r \,\&\, E_w)$ where

  E_r= ”A and B use the same IP range”

  E_w= ”A and B have both edited this wiki”

- Assume for the moment that E_r and E_w are **statistically independent**, i.e. they don't influence each other's probabilities:

  $P(E_r|E_w) = P(E_r)$ and $P(E_w|E_r) = P(E_w)$

  Then $P(E_r \,\&\, E_w) = P(E_r)\, P(E_w)$.

- Guessing $P(H)=10^{-9}$ and $P(E_r)=2^{-18}$ as previously, and $P(E_w) = 0.001$ (i.e. one million surfers have edited this wiki), Bayes would give

  $P(H|E_r \,\&\, E_w) = P(H|E_r)/P(E_w) \approx 0.026\% / 0.001 = 26\%$

Combining evidence: Problems

- NB: In reality E_r (using the IP range of A) and E_w (having edited the wiki) will not be entirely independent:
  - Extreme example: Only one person (A) has ever edited the wiki. Then $P(E_r|E_w) = 1$, which is certainly not equal to $P(E_r)$ unless that IP range is the whole Internet (i.e. no CheckUser evidence is present).
  - More realistic: The language of a project certainly influences $P(E_r)$. For example, on the German Wikipedia, ISPs from Germany, Austria and Switzerland are over-represented.
- One possibility to estimate $P(E_r \& E_w)$ instead: Look how frequently the range occurs in the recent changes of that wiki

Prosecutor's fallacy

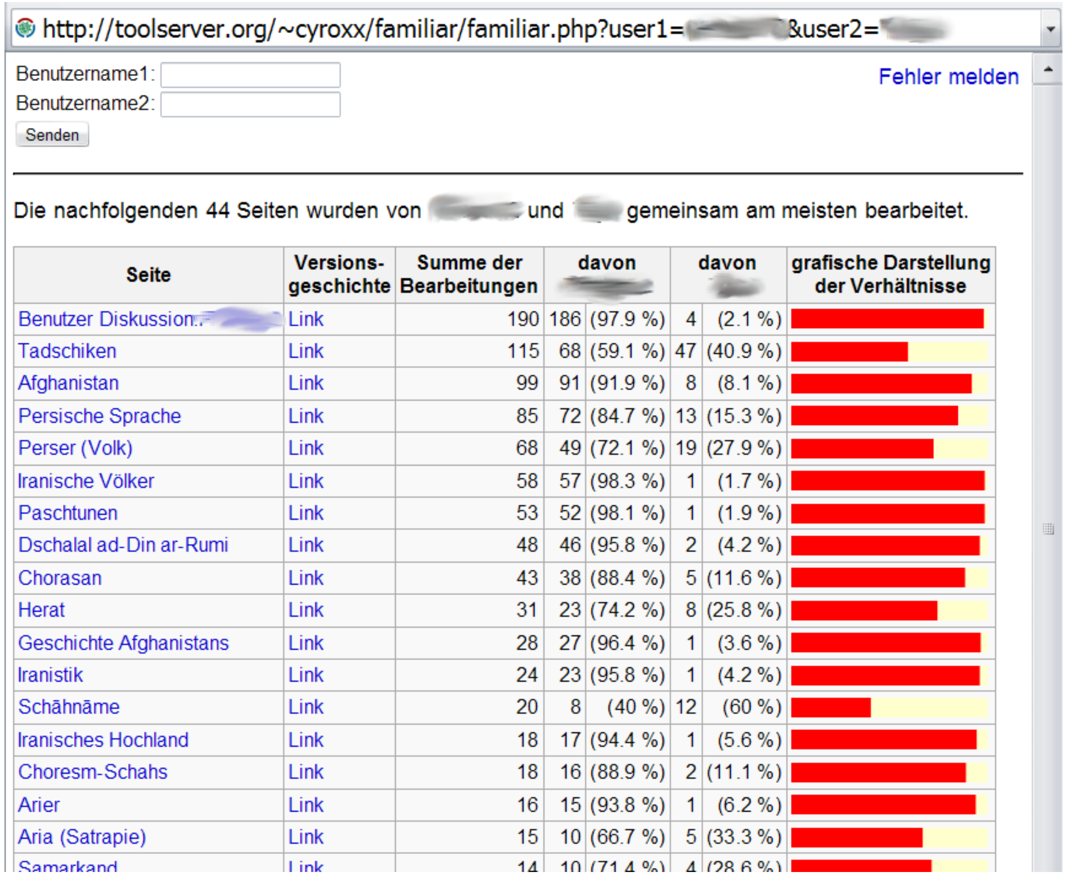
- “Fishing for socks”: Look for B's which share A's IP range. Then argue
  - *“The probability for B using the same range as A by pure coincidence is really low, so it is very unlikely that B is not a sockpuppet of A”*
- Fallacy: First part is true (remember $P(E_r)=2^{-18}$), but B was specifically selected for this property, not by a random process (“pure coincidence”).
- Known as “prosecutors's fallacy” for its occurrence in several real-life court cases

Combining with non-CU evidence, defendant's fallacy

- Recall that in our numerical example, P(H|E) was still small (nowhere near 1), even when combining IP range and being an editor
- Other evidence from CU: User agents and temporal patterns (e.g. A uses an IP at 12:07 and 12:20 pm, and B the same IP at 12:12 pm). Sometimes sufficient to conclude sockpuppetry, but:
- Usually, the CU output has to be complemented by other evidence to reach a sound conclusion.
- “Defendant's fallacy”: B argues “Tens of thousands of other people use this IP range besides me and A. So P(A=B) < 0.01%.” - Ignores that other evidence may be present.

Similar interests

- Just a few personal interests and cultural preferences can suffice to identify an individual (e.g. Narayanan,Shmatikov: *How to Break Anonymity of the Netflix Prize Dataset,* 2007)
- Frequent argument in sockpuppet cases on Wikipedia: “Both accounts edit articles about (special topic X) and (unrelated special topic Y)”
- Afaik no systematic analysis yet. But there is a tool which, for two users, displays articles that both have edited



http://toolserver.org/~cyroxx/familiar/familiar.php?user1=[redacted]&user2=[redacted]

Benutzername1:

Benutzername2:

Senden

Fehler melden

Die nachfolgenden 44 Seiten wurden von [redacted] und [redacted] gemeinsam am meisten bearbeitet.

Seite	Versions-geschichte	Summe der Bearbeitungen	davon [redacted]		davon [redacted]		grafische Darstellung der Verhältnisse
Benutzer Diskussion:[redacted]	Link	190	186	(97.9 %)	4	(2.1 %)	
Tadschiken	Link	115	68	(59.1 %)	47	(40.9 %)	
Afghanistan	Link	99	91	(91.9 %)	8	(8.1 %)	
Persische Sprache	Link	85	72	(84.7 %)	13	(15.3 %)	
Perser (Volk)	Link	68	49	(72.1 %)	19	(27.9 %)	
Iranische Völker	Link	58	57	(98.3 %)	1	(1.7 %)	
Paschtunen	Link	53	52	(98.1 %)	1	(1.9 %)	
Dschalal ad-Din ar-Rumi	Link	48	46	(95.8 %)	2	(4.2 %)	
Chorasan	Link	43	38	(88.4 %)	5	(11.6 %)	
Herat	Link	31	23	(74.2 %)	8	(25.8 %)	
Geschichte Afghanistans	Link	28	27	(96.4 %)	1	(3.6 %)	
Iranistik	Link	24	23	(95.8 %)	1	(4.2 %)	
Schāhnāme	Link	20	8	(40 %)	12	(60 %)	
Iranisches Hochland	Link	18	17	(94.4 %)	1	(5.6 %)	
Choresm-Schahs	Link	18	16	(88.9 %)	2	(11.1 %)	
Arier	Link	16	15	(93.8 %)	1	(6.2 %)	
Aria (Satrapie)	Link	15	10	(66.7 %)	5	(33.3 %)	
Samarkand	Link	14	10	(71.4 %)	4	(28.6 %)	

Style analysis

- Users try to find significant similarities in the language used by suspected sockpuppets, such as repeated unusual typos, peculiar abbreviations, punctuation habits etc.

Stylometry, forensic linguistics

- History: Attempts to determine authorship of Shakespeare's works, the Federalist papers, the Unabomber manifesto...
- Underlying assumption: While people vary their writing style according to occasion, genre, mood etc., there exist persistent habits and traits which distinguish individual writers.
- Which properties can be regarded as persistent is often controversial, but many successes
- How does it work? Example: “tf-idf similarity”

tf-idf similarity

- In a collection (*corpus*) of texts (*documents)* d, each consisting of words (*terms*) t:
- The tf-idf *weight* (term frequency-inverse document frequency) of a term t measures its importance within a document d, relative to its importance in the whole corpus. (Exact definition varies)
- tf-idf weight (of t in d) = tf · idf, where:
  - tf = *term frequency* of t in d. This is the number of occurrences of t in d, divided by the overall number of words in d.
  - idf = *inverse document frequency* of t in the corpus. This is the logarithm of the quotient of the number of all documents divided by the number of documents where t occurs
- If t1 and t2 have the same frequency in d, but t1 is unusual in other documents while t2 is equally common in most other documents (e.g. t2="and" in English texts), then tf-idf(t1,d) > tf-idf(t2,d)

tf-idf similarity

- Listing the tf-idf weights of all terms t for one d gives a vector. Angle between two of these vectors is a measure of similarity between the two documents, regarding word usage.
- Now combine the text contributions (or the edit summaries) of an user account into a document d, and take the contributions of all accounts on the wiki as the corpus. The tf-idf vector of d says something about the vocabulary preferences of that account. Accounts with a higher tf-idf similarity are more likely to be sockpuppets of the same person.

tf-idf and other similarity measures as sockpuppet evidence

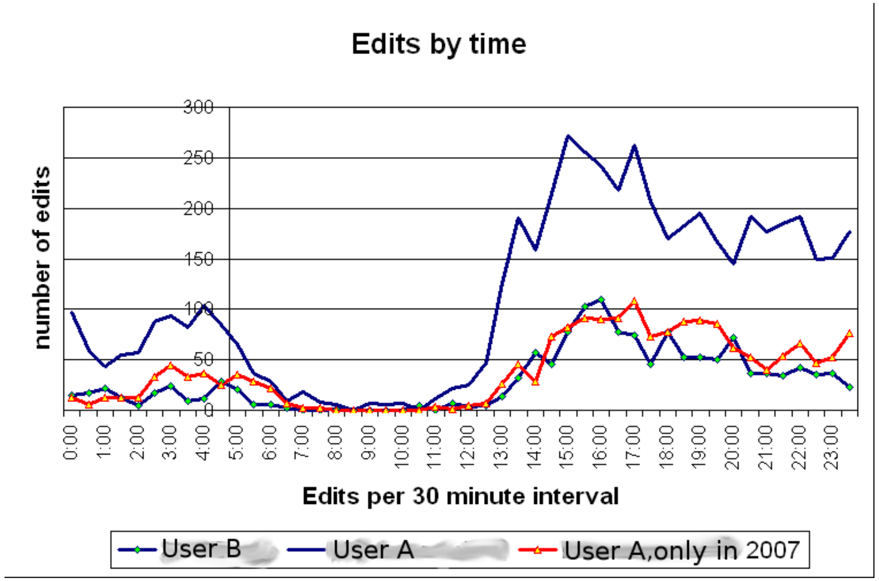
- Novok, Raghavan, Tomkins (*Anti-Aliasing on the Web*, 2004) evaluated tf-idf and other similarity measures on a corpus of postings of the Courttv.com webforum, concluding
  - “matching aliases to authors with accuracy in excess of 90% is practically feasible in online environments”
- tf-idf similarity was for a sockpuppet investigation on the English Wikipedia in 2008 (by User:Alanyst):
  - Corpus = aggregated edit summaries of all users which had between 500 and 3500 edits in 2007 (11,377 accounts). All users/all years would have been to computationally expensive.
  - To improve independence, manually excluded terms specific to the topic that the suspected sockpuppets were editing
  - Account B came out closest to A, and account A 188th closest to B (among the 11,377 tested accounts)

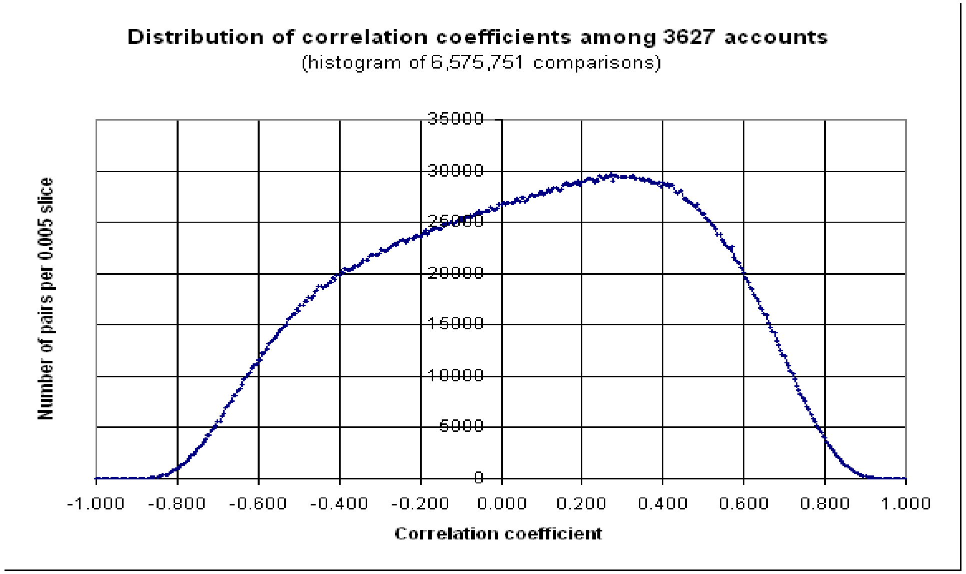
Selection bias

- Fallacy: From many evidence parameters E select a “nice one” where A and B match (i.e. silently discard the others where they don't match): not the same P as if parameters were chosen independently of the outcome
- Example: Lincoln-Kennedy coincidences
  - Both presidents were shot on a Friday!
  - Both were elected to the congress in '46 !
  - Both were elected to the presidency in '60 !
  - Both surnames have 7 letters!

Temporal editing patterns

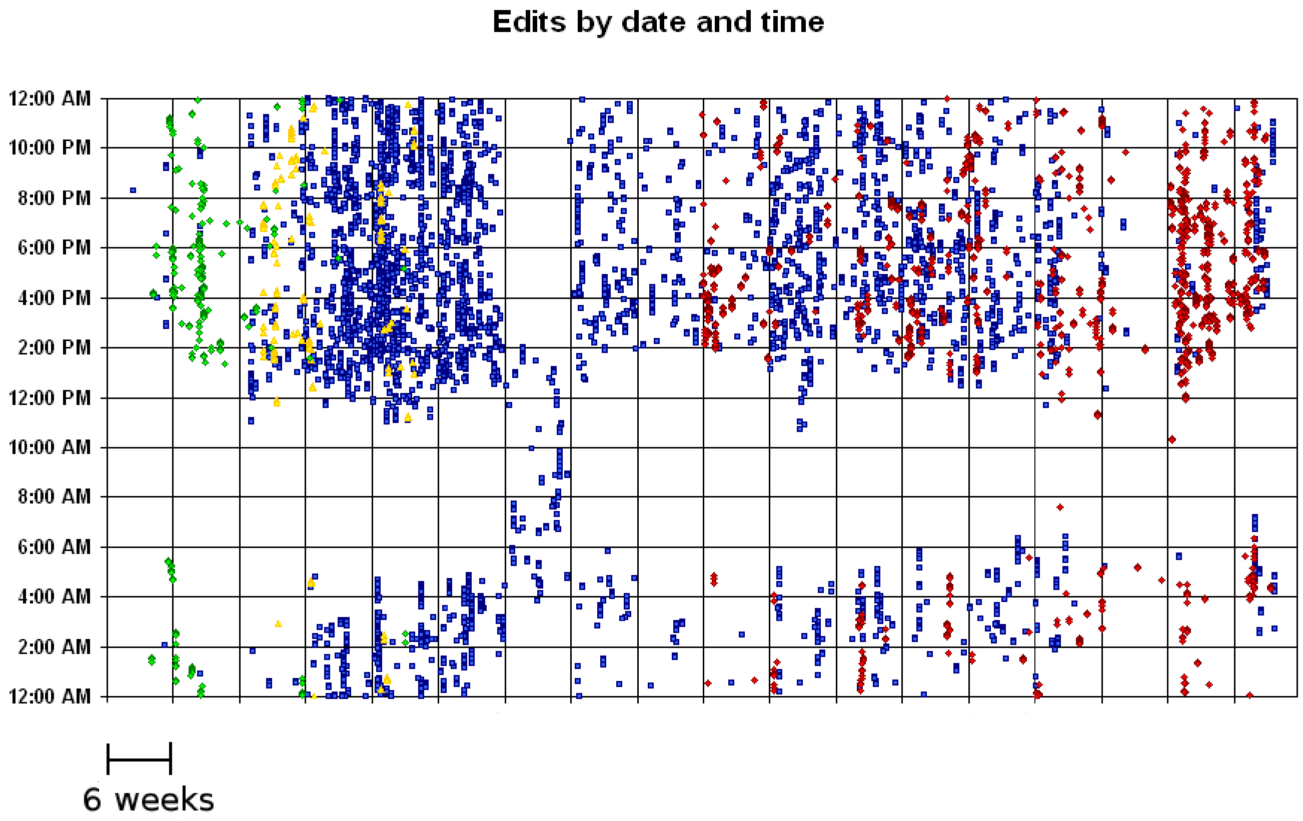
- Count edit frequency over time of day
- Compute correlation coefficient between the curves for A and B
- Histogram of correlation coefficients gives an estimate for P(E). Done by User:Cool Hand Luke on the English Wikipedia for 3627 accounts:



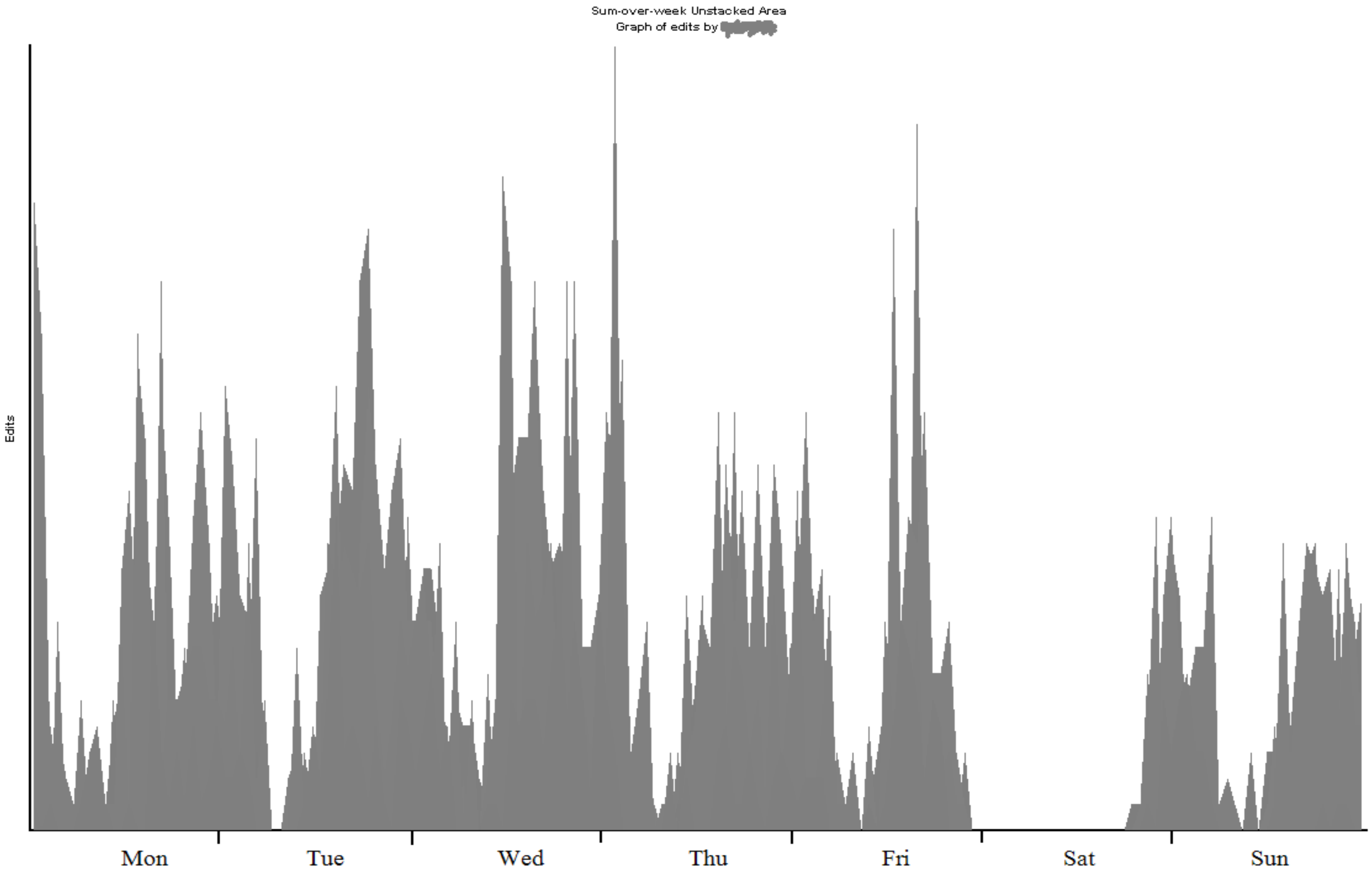


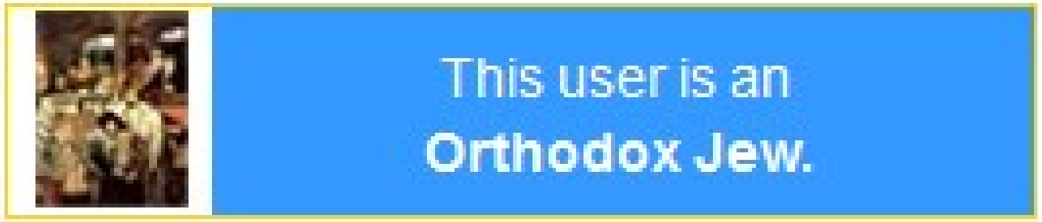
Temporal editing patterns & real life info

- Case from the English Wikipedia: A certain person is suspected to edit under certain accounts. From public statements, it is known that this person usually lives on the US East Coast, but spent some weeks in India around a certain date.



Quiz question: What can one say about this user ?




This user is an **Orthodox Jew.**

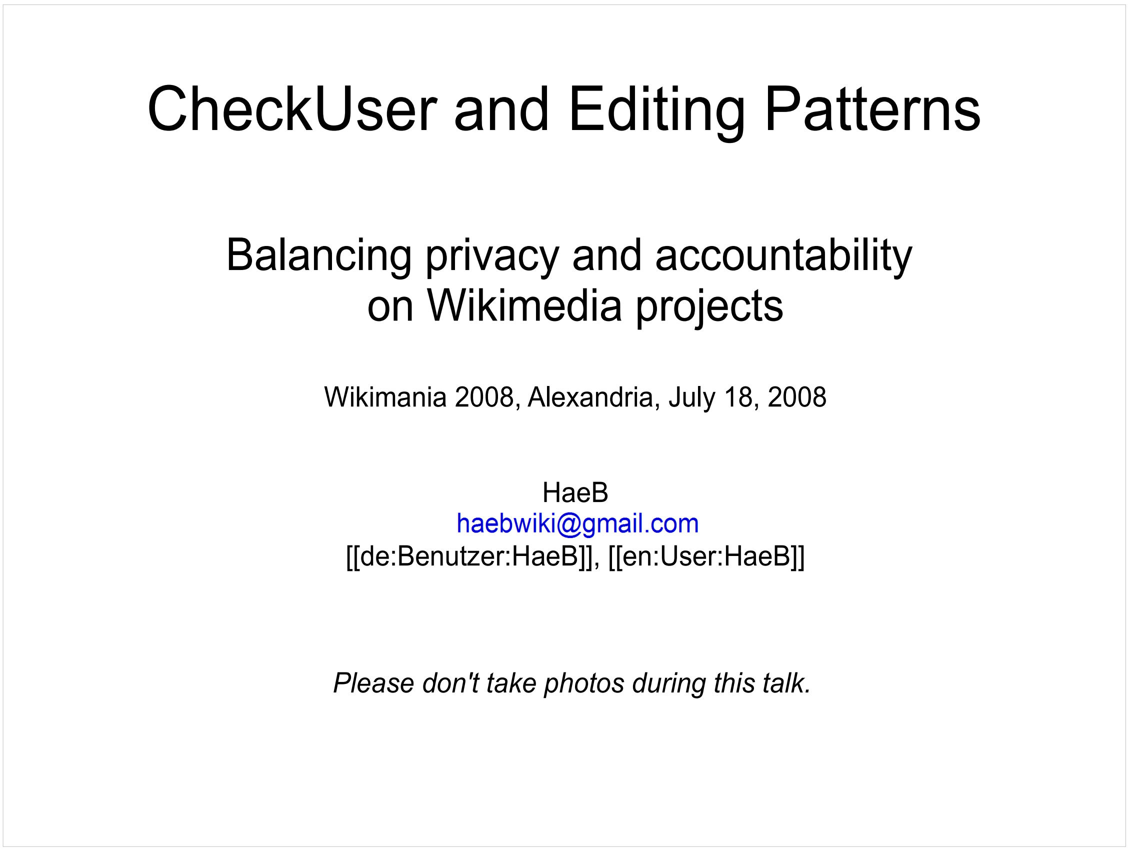
- Ramadan, Sunday morning church attendance, fan of a weekly TV show .... Analysis of temporal patterns gives rise to many more such privacy concerns.
- The first online tool which generated such weekly graphs was made opt-in after privacy concerns were raised.

Privacy laws and Checkuser

- If personal data is stored, European privacy laws such as the German Bundesdatenschutzgesetz
  - Require a purpose for storage (not just “because we have it”)
  - Prohibit revealing the data to third parties unless subject agreed to
- Privacy policy seems to address these
- (Old) privacy policy mainly governs the public release of CheckUser data, but many users feel their privacy would be violated by unwarranted checks too
- Sockpuppet confirmation can mean a privacy violation too
- Not sure about data aggregation (recall the weekly graph tool)
- IANAL

Conclusions

- The community has developed a lot of expertise, some homegrown tools and clever techniques to generate sockpuppet evidence.
- There is still much more potential for more formalized and more automated analysis, applying research from several fields.
- Wikipedia contributors don't just give their time to the project, but pay with their privacy, too.



CheckUser and Editing Patterns

Balancing privacy and accountability
on Wikimedia projects

Wikimania 2008, Alexandria, July 18, 2008

HaeB
haebwiki@gmail.com
[[de:Benutzer:HaeB]], [[en:User:HaeB]]

Please don't take photos during this talk.

This PDF contains the slides as they were used in the talk, some notes (mainly remarks made verbally during the talk) and a few references.

Due to time restrictions (the talk lasted about 35 minutes, including audience questions, instead of the scheduled 45 minutes), some of these slides had to be skipped in Alexandria.

Many thanks to: Bdk, Elian, Hei_ber

What are sockpuppets?

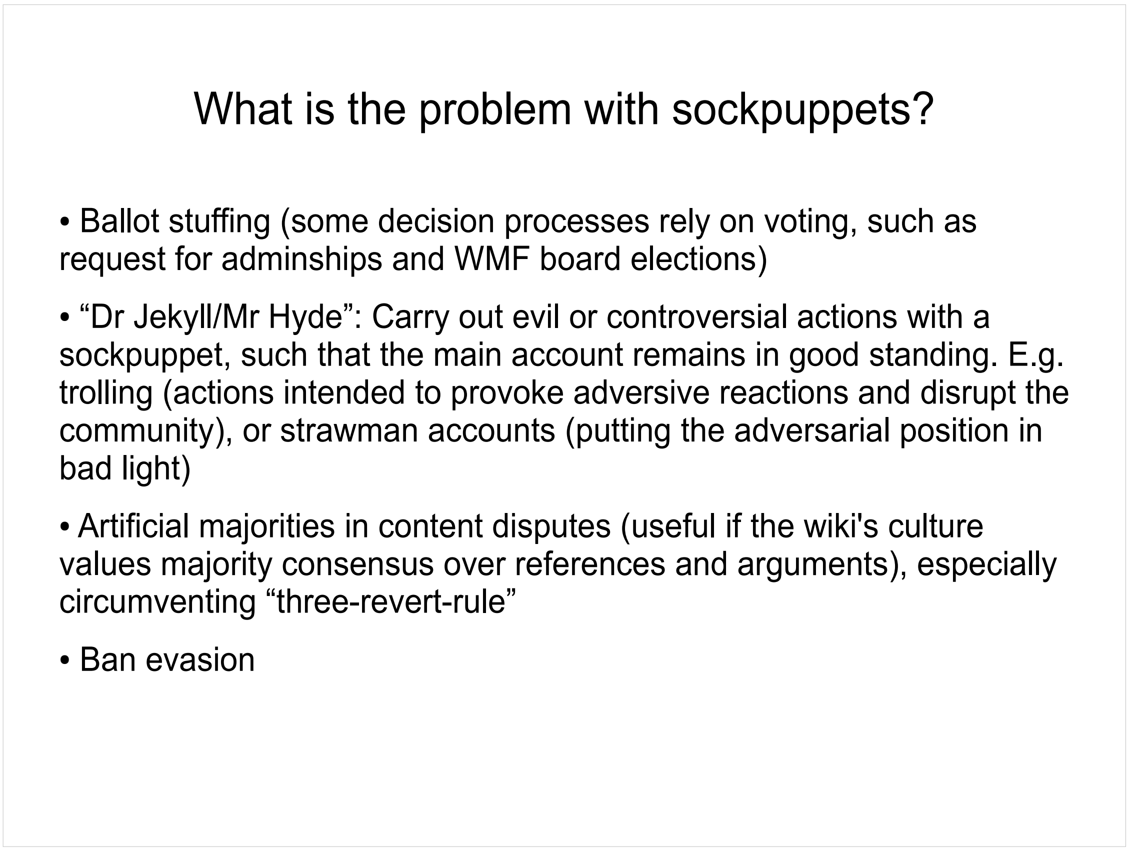
- Wikipedia and sister projects rely on being open: Anyone can edit, anyone can get an account.
- No ID control when registering (not even email address required)
- Many legitimate uses for multiple accounts
- “Sockpuppet” often implies deceptive intention, think ventriloquist

First explain the problem that CU is intended to solve

For a sample of Wikipedia's sock puppet folklore, visit Phoebe's workshop at this conference

Examples of legitimate uses:
- protect login data when accessing over insecure connection (open WLAN)
- protect real-life privacy
- avoid real-life harassment

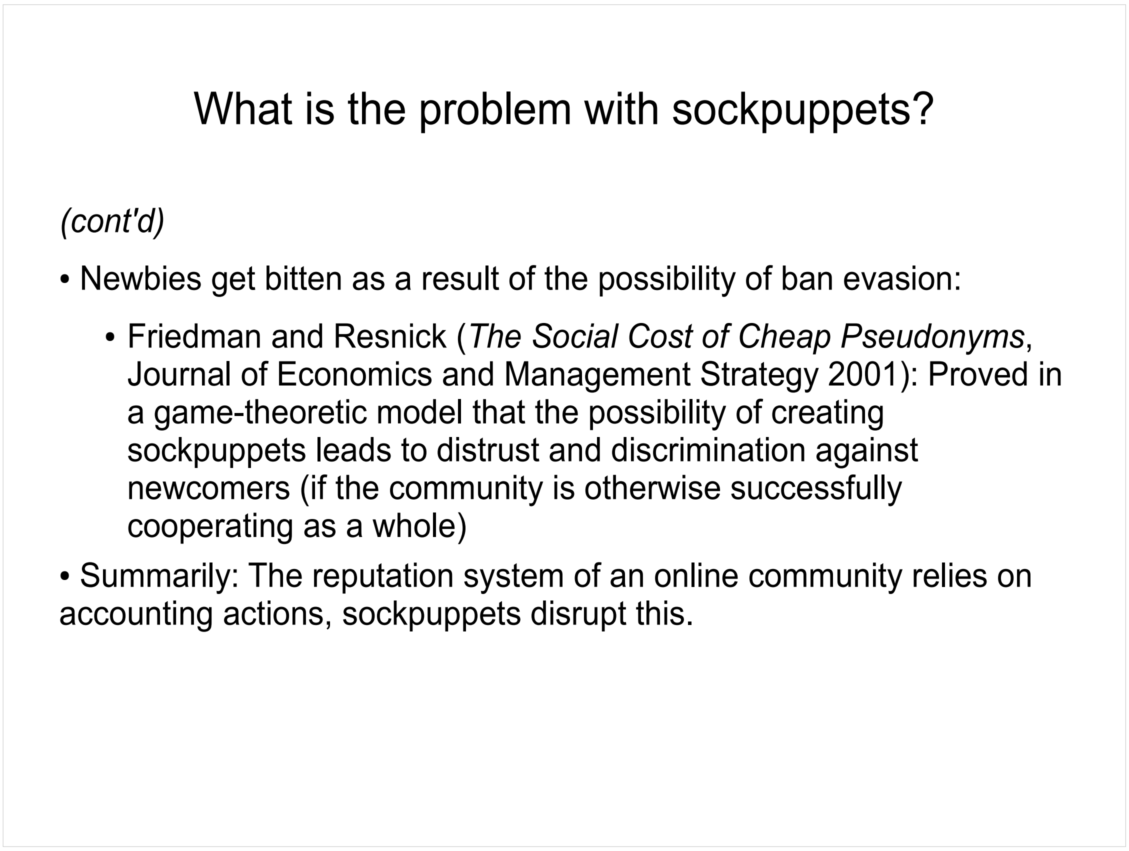
Ventriloquist photo from
http://commons.wikimedia.org/wiki/Image:Mallory_Lewis_and_Lamb_Chop.jpg
(Public Domain)



What is the problem with sockpuppets?

- Ballot stuffing (some decision processes rely on voting, such as request for adminships and WMF board elections)
- “Dr Jekyll/Mr Hyde”: Carry out evil or controversial actions with a sockpuppet, such that the main account remains in good standing. E.g. trolling (actions intended to provoke adversive reactions and disrupt the community), or strawman accounts (putting the adversarial position in bad light)
- Artificial majorities in content disputes (useful if the wiki's culture values majority consensus over references and arguments), especially circumventing “three-revert-rule”
- Ban evasion

- Voting is discouraged in principle, but ...
- Real life strawman example: On de:WP, a long time right-wing sockpuppeteer sometimes creates “leftist” sockpuppets.
- In the abstract theory of social networks and reputation systems, ballot stuffing/creating artificial majorities is known as a “sybil attack”.



What is the problem with sockpuppets?

(cont'd)

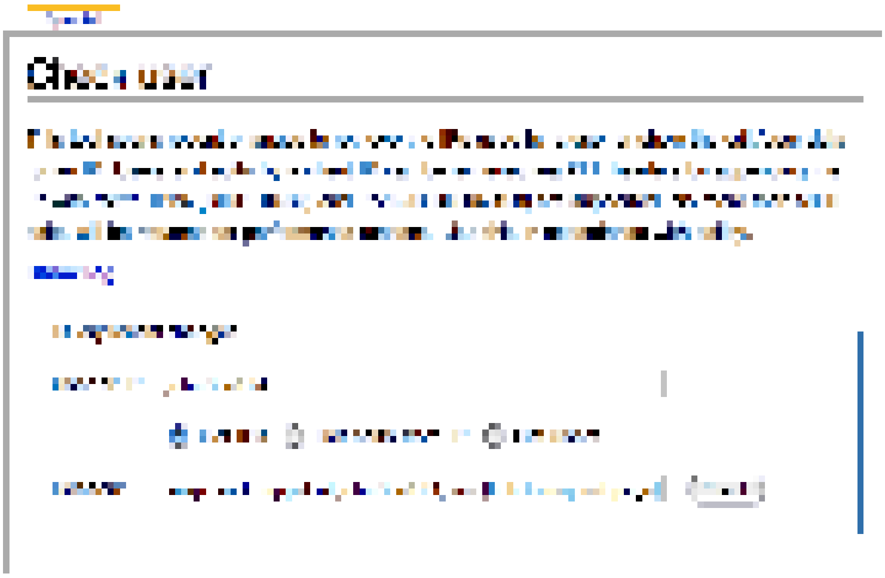
- Newbies get bitten as a result of the possibility of ban evasion:
  - Friedman and Resnick (*The Social Cost of Cheap Pseudonyms*, Journal of Economics and Management Strategy 2001): Proved in a game-theoretic model that the possibility of creating sockpuppets leads to distrust and discrimination against newcomers (if the community is otherwise successfully cooperating as a whole)
- Summarily: The reputation system of an online community relies on accounting actions, sockpuppets disrupt this.

(Friedman and Resnick assume that a few malicious players are always present.)

Example of an online community (almost) without accounts, therefore without reputation system: 4chan

Friedman, E. and P. Resnick (2001). "The Social Cost of Cheap Pseudonyms." **Journal of Economics and Management Strategy** 10(2): 173-199.
Preprint available at
http://www.si.umich.edu/~presnick/papers/identifiers/index.html

The CheckUser tool in MediaWiki



“Get IPs”: Retrieve IP addresses from which this user account has edited

“Get edits from IP”: Retrieve edits (logged-in or not) which have been made from this IP

“Get users”: List accounts which have made edits from this IP

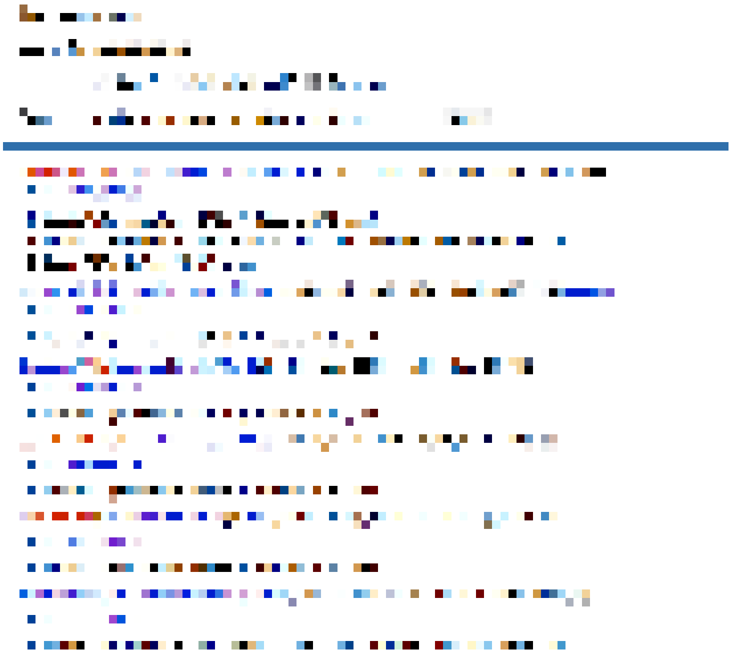
Screenshot provided by Bdk, http://commons.wikimedia.org/wiki/Image:CheckUser1.png (version of 19:02, 17 July 2008), Public Domain

Get IPs of a logged-in user

Note: The IPs in this mock-up example don't actually exist. But it intends to demonstrate a common real-life phenomenon: A user edits from changing (“dynamic”) IPs, but they always stay in the same “range” (here: 222.333.444.XYZ).

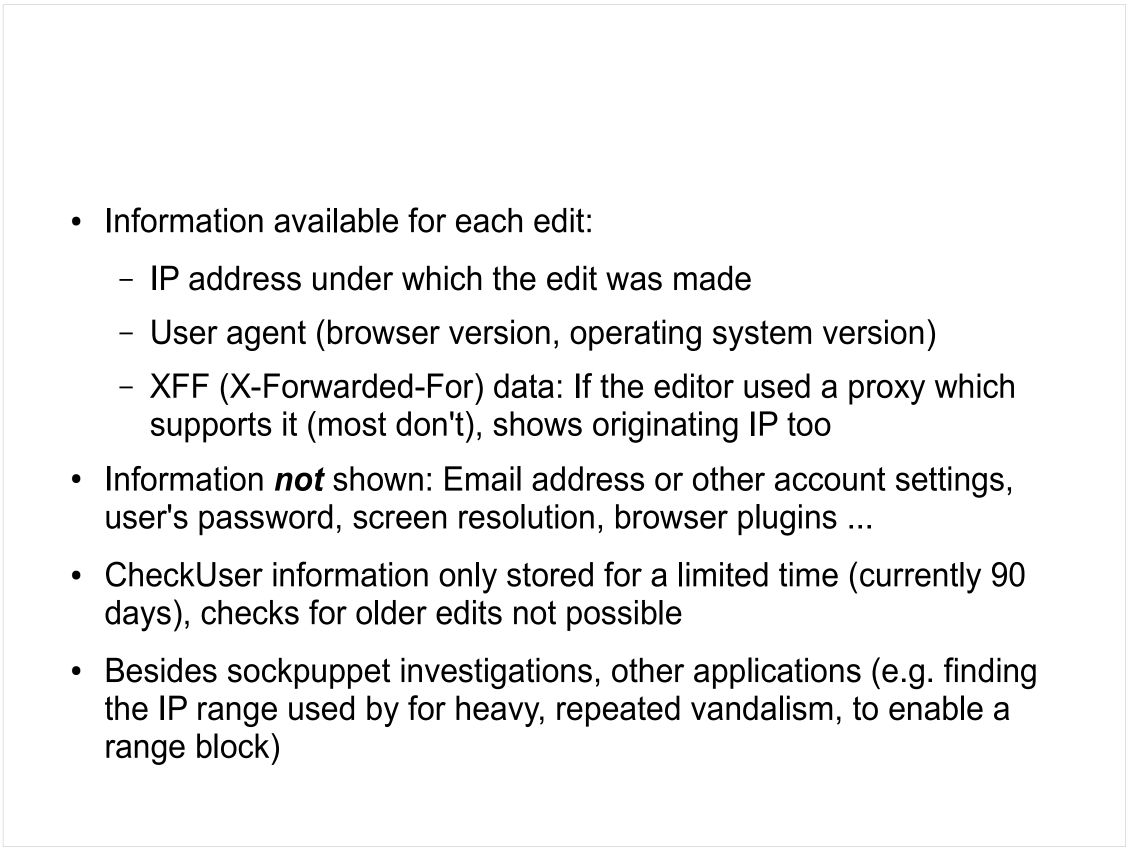
Screenshot provided by Bdk, http://meta.wikimedia.org/wiki/Image:CheckUser3.png (version of 17:37, 24 May 2007), Public Domain

Get edits



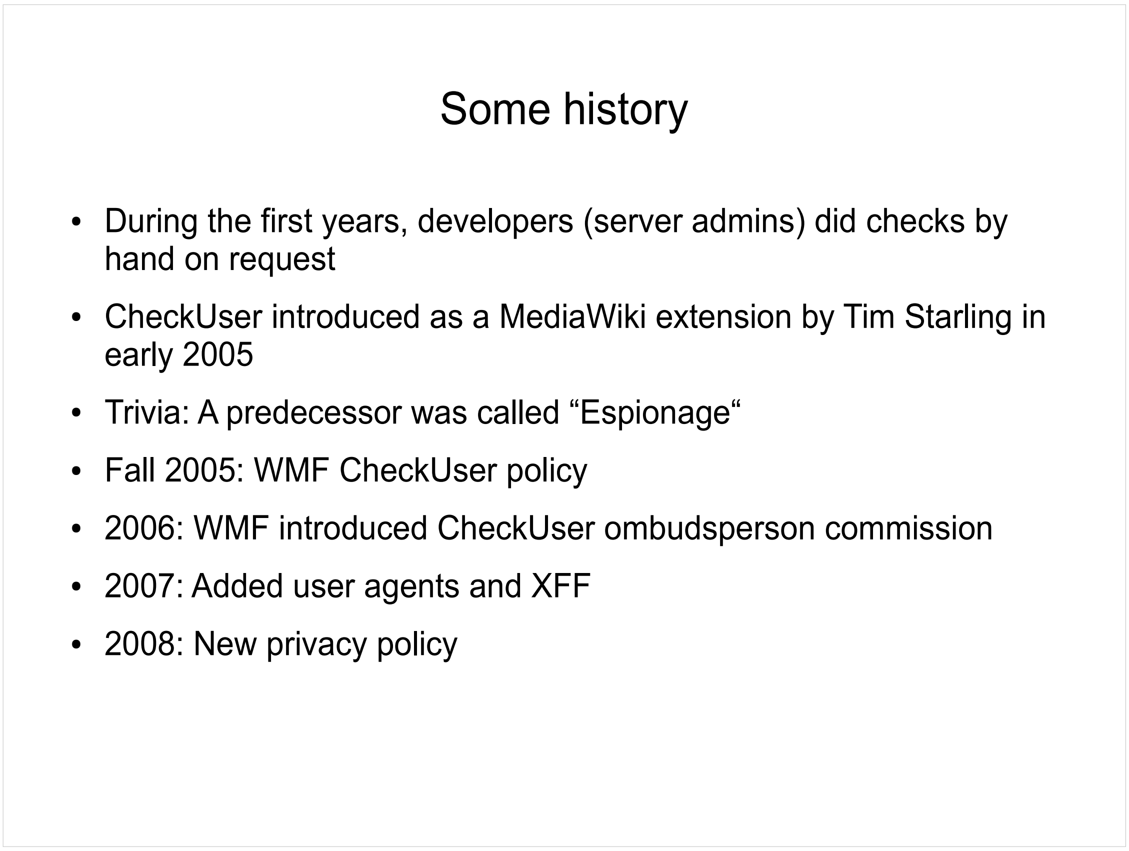
Accounts editing from the same IP might not necessarily belong to the same user. But here, the similarity of the usernames (i.e. additional evidence which is independent of the CheckUser data) allows to conclude with some certainty that the first four accounts are sockpuppets. Not so for the last account; which also has a different user agent string, i.e. seems to use a different browser. (However, this is no proof of Flowerhunter's “innocence” either, since a sockpuppeteer can easily change between browsers, or even forge the user agent string.)

Screenshot provided by Bdk, http://commons.wikimedia.org/wiki/Image:CheckUser2.png (version of 18:38, 17 July 2008), Public Domain



- Information available for each edit:
  - IP address under which the edit was made
  - User agent (browser version, operating system version)
  - XFF (X-Forwarded-For) data: If the editor used a proxy which supports it (most don't), shows originating IP too
- Information ***not*** shown: Email address or other account settings, user's password, screen resolution, browser plugins ...
- CheckUser information only stored for a limited time (currently 90 days), checks for older edits not possible
- Besides sockpuppet investigations, other applications (e.g. finding the IP range used by for heavy, repeated vandalism, to enable a range block)

(Some browser add-ons do show up in the user agent, though.)



Some history

- During the first years, developers (server admins) did checks by hand on request
- CheckUser introduced as a MediaWiki extension by Tim Starling in early 2005
- Trivia: A predecessor was called “Espionage“
- Fall 2005: WMF CheckUser policy
- 2006: WMF introduced CheckUser ombudsperson commission
- 2007: Added user agents and XFF
- 2008: New privacy policy

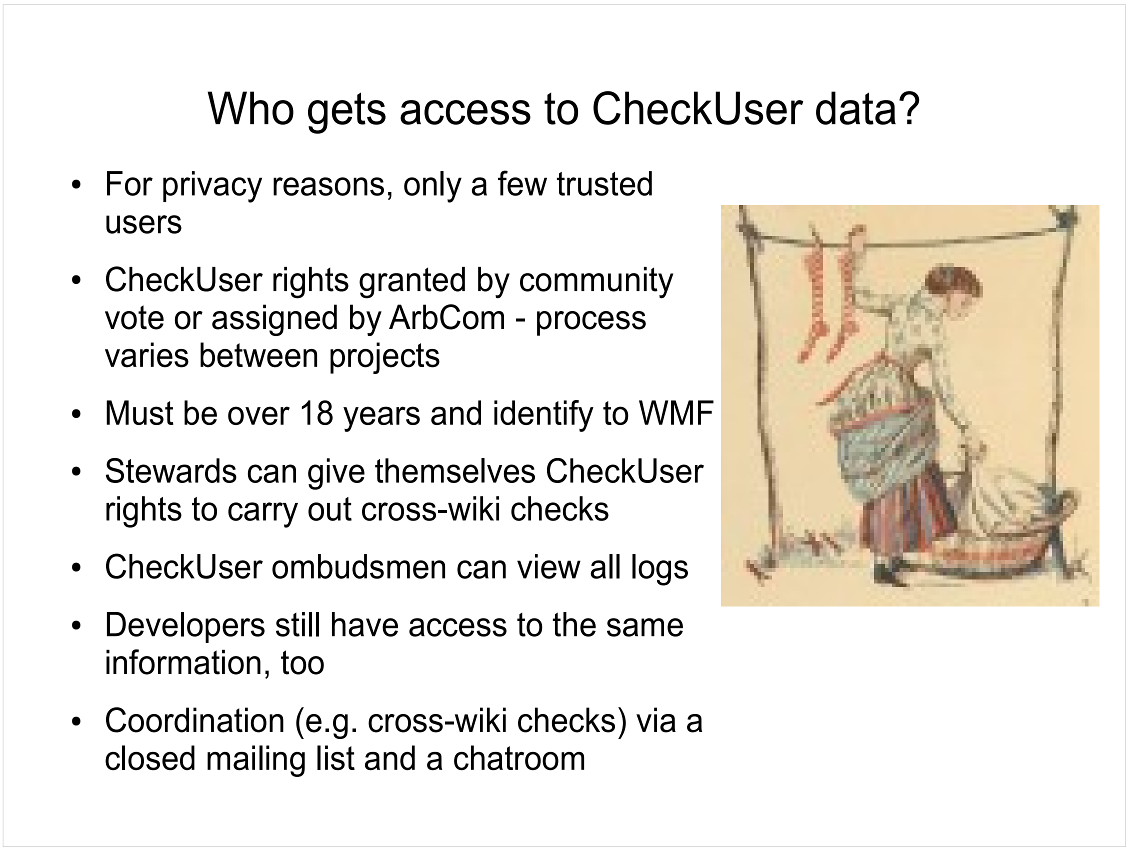
It was decided to separate server access from access to logged IPs.

In early 2005, just one Checkuser at the English WP

http://meta.wikimedia.org/wiki/CheckUser_policy

2007: Access to non-public data resolution (requires persons with CheckUser rights to be at least 18 and to identify themselves to the WMF)

http://wikimediafoundation.org/wiki/Privacy_policy



Who gets access to CheckUser data?

- For privacy reasons, only a few trusted users
- CheckUser rights granted by community vote or assigned by ArbCom - process varies between projects
- Must be over 18 years and identify to WMF
- Stewards can give themselves CheckUser rights to carry out cross-wiki checks
- CheckUser ombudsmen can view all logs
- Developers still have access to the same information, too
- Coordination (e.g. cross-wiki checks) via a closed mailing list and a chatroom

Policy requires each project to have at least two (or none) CheckUsers, to ensure mutual control

Toolserver root admins (ca. 5 people) can also access IPs of logged-in editors on all projects

Cross-wiki checks: Useful mainly to find sleeper accounts (created to carry out vandalism edits which require auto-confirmation), and accounts created for harrassment by choice of the user name

Image: Illustration to “Sing a Song of Sixpence” by Walter Crane, available at

http://commons.wikimedia.org/wiki/Image:Sing_a_sing_of_sixpence_-_illustration_by_Walter_Crane_-_Project_Gutenberg_eText_18344.jpg

(Public Domain), discovered by User Hozro of the German Wikipedia. Used here in analogy to http://commons.wikimedia.org/wiki/Image:Admin_mop.PNG

CheckUser log

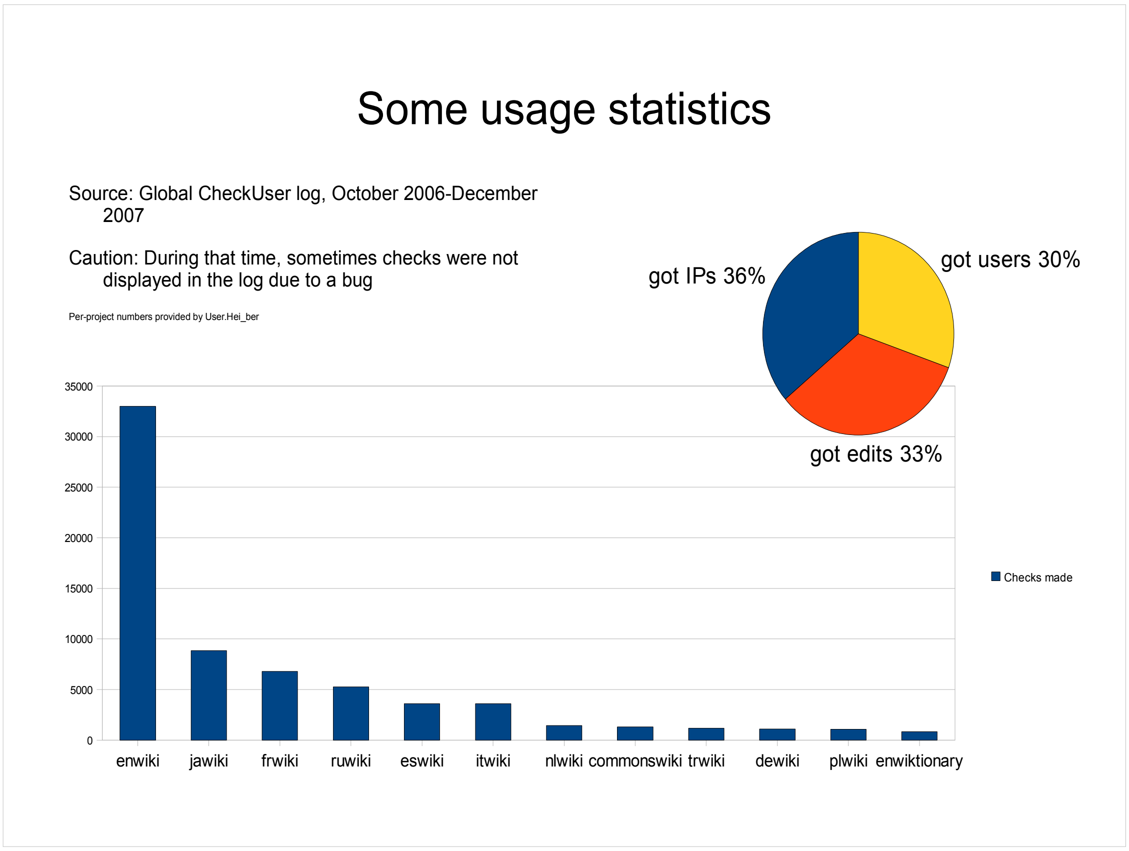
You see the date of each check, who did it, which account/IP was checked, and the reason given. ("WP:CU/A" is referring to the CheckUser request page http://de.wikipedia.org/wiki/WP:CU/A)

The results themselves are not visible in the log, but can often be guessed from it (in this mockup example, one would assume that 333.222.111.86 was checked because this IP turned out to be used by user "Troll ABC")

Screenshot provided by Bdk, http://commons.wikimedia.org/wiki/Image:CheckUser_log.png (version of 15:10, 17 July 2008), Public Domain

CheckUser log

- Visible to other CheckUsers on the same projects, CheckUser ombudsmen and developers
- Used to be visible to CheckUsers from all projects (global log, has been disabled for some time due to technical problems)



Some usage statistics

Source: Global CheckUser log, October 2006-December 2007

Caution: During that time, sometimes checks were not displayed in the log due to a bug

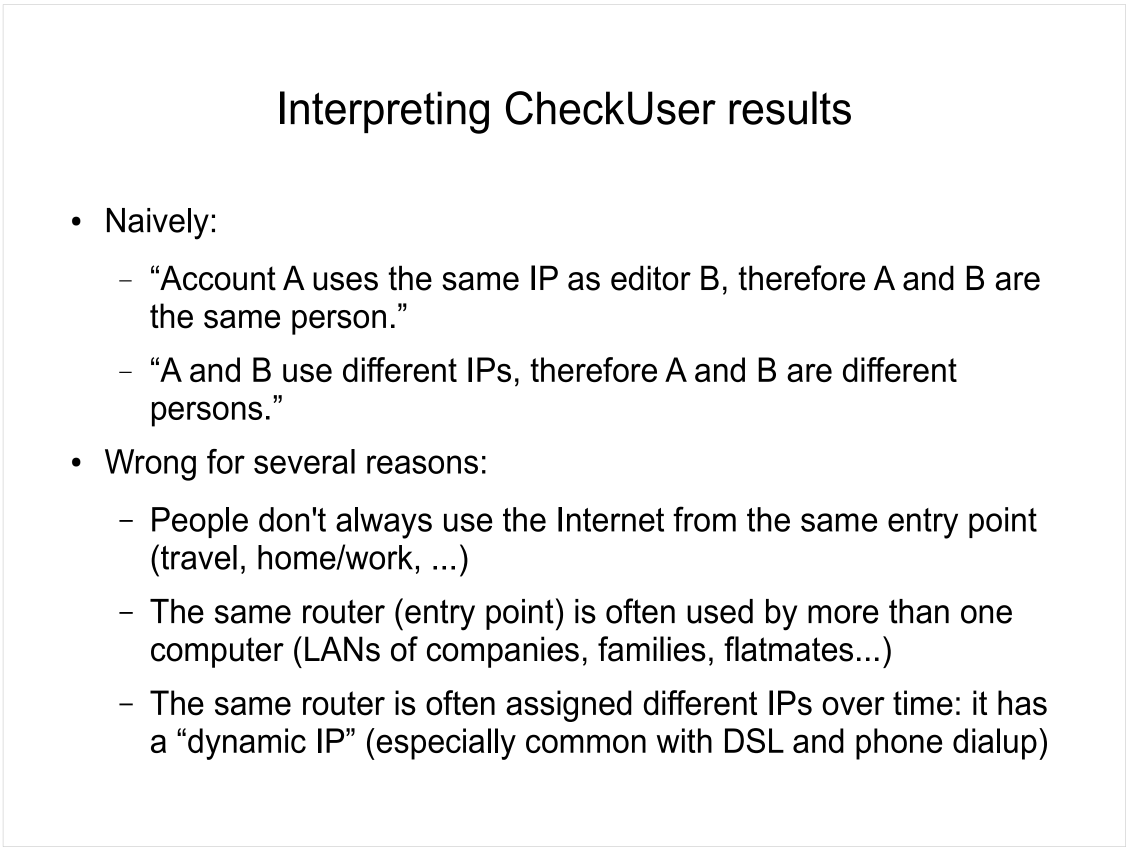
Per-project numbers provided by User.Hei_ber

Possible reasons for the high number of checks on the English Wikipedia? (question raised during the talk)

First note: These are absolute numbers, not accounting for the size of each project. The number of checks per edit on each project would be more significant, but unfortunately edit stats are not available for en: after October 06.

Some guesses (Disclaimer: I am only familiar with de: apart from en:):

- en: might rely more heavily on certain formal rules like 3RR and “Votes” for deletion which create a higher incentive to use sockpuppets
- On de:, admins might have more leeway to block suspected sockpuppets passing the “duck test” without CU evidence.
- Different CU policies on the projects (cf. http://meta.wikimedia.org/wiki/CheckUser_policy/Local_policies), e.g. on some projects CU is only done on request, on others CU start checks on their own initiative
- Different privacy expectations by the communities: Is a check only accepted in cases of grave abuse, or also in cases of less serious policy violations? Do users mind being checked if the results are not published?
- Maybe also shaped by different privacy cultures in different countries (e.g. US vs. EU)

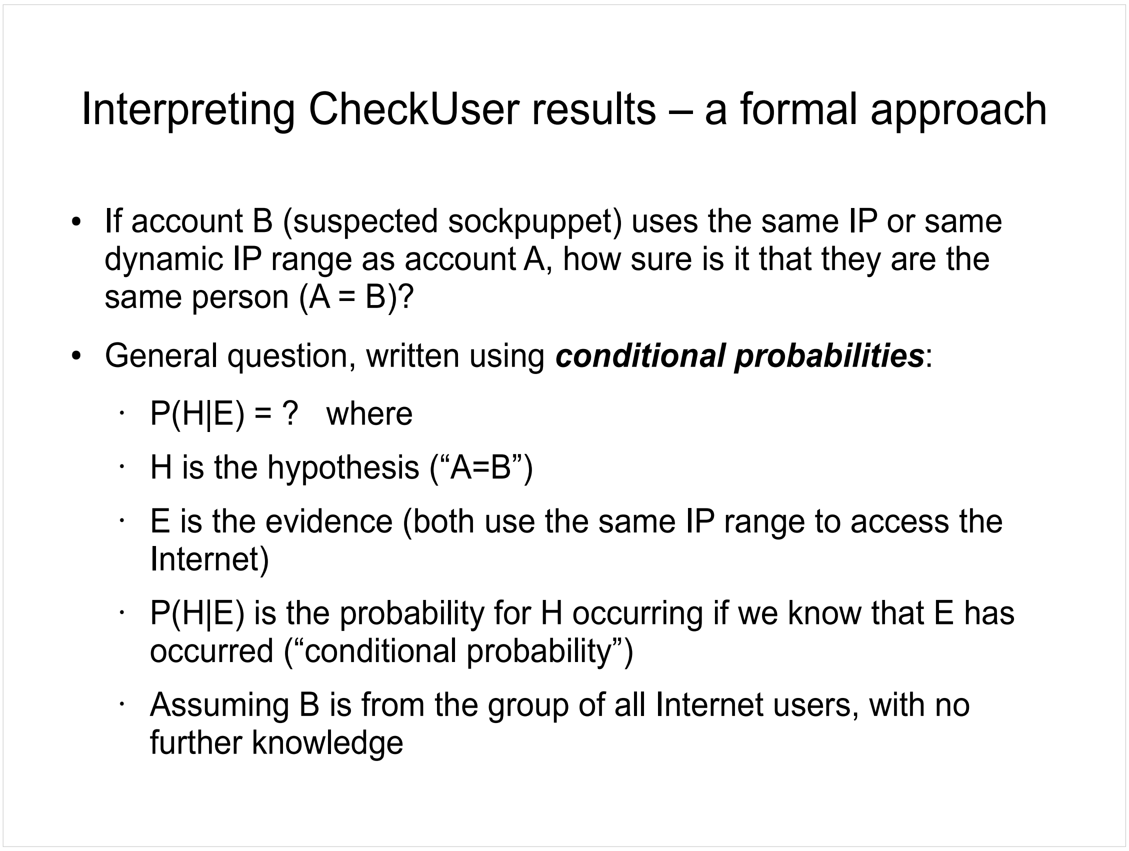


Interpreting CheckUser results

- Naively:
  - “Account A uses the same IP as editor B, therefore A and B are the same person.”
  - “A and B use different IPs, therefore A and B are different persons.”
- Wrong for several reasons:
  - People don't always use the Internet from the same entry point (travel, home/work, ...)
  - The same router (entry point) is often used by more than one computer (LANs of companies, families, flatmates...)
  - The same router is often assigned different IPs over time: it has a “dynamic IP” (especially common with DSL and phone dialup)

On this and the following slides, user agents and XFF are ignored for simplicity. (And of course not every computer which is assigned an IP address in the Internet is a router; the above just refers to a typical situation.)

As long as the entry point stays physically the same, its dynamic IPs usually still fall within the same subnet or an even narrower IP range (cf. RFC 1518: “the assignment of IP addresses must be ... consistent with the actual physical topology of the Internet”).

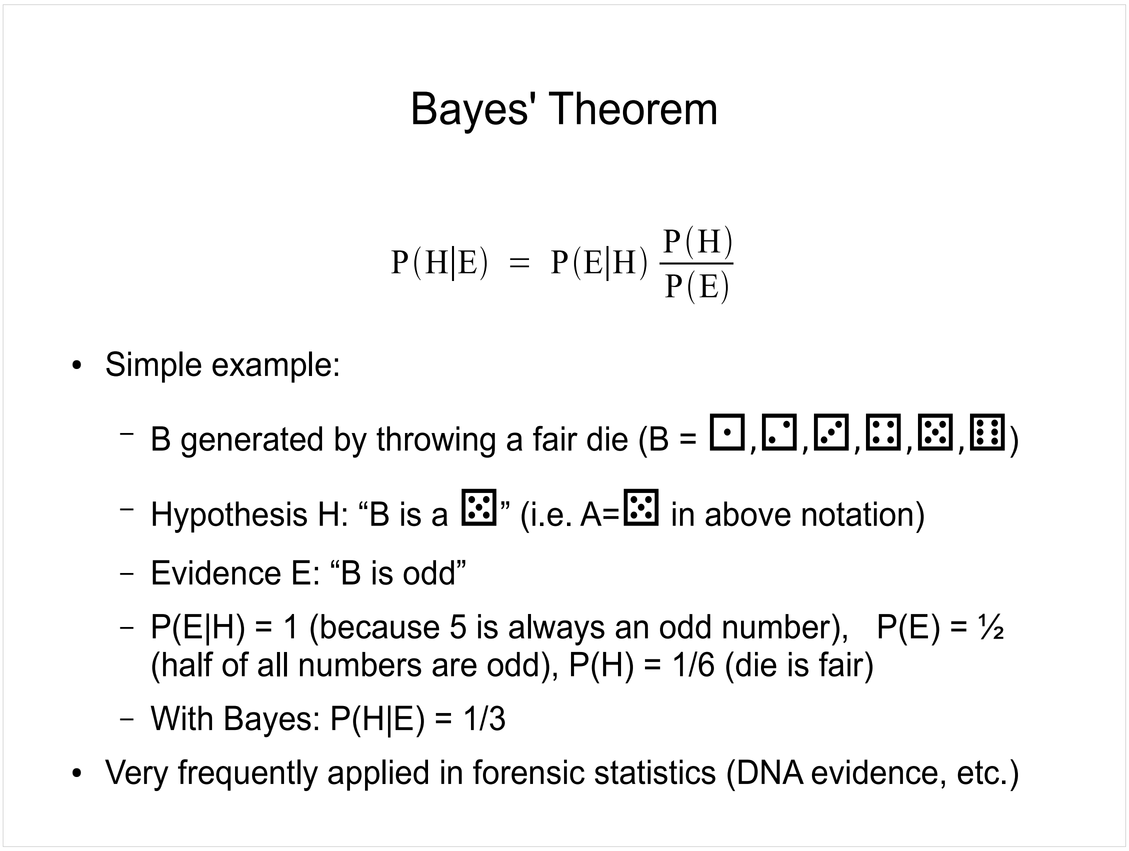


Interpreting CheckUser results – a formal approach

- If account B (suspected sockpuppet) uses the same IP or same dynamic IP range as account A, how sure is it that they are the same person (A = B)?
- General question, written using ***conditional probabilities***:
  - P(H|E) = ?   where
  - H is the hypothesis (“A=B”)
  - E is the evidence (both use the same IP range to access the Internet)
  - P(H|E) is the probability for H occurring if we know that E has occurred (“conditional probability”)
  - Assuming B is from the group of all Internet users, with no further knowledge

With experience, CheckUsers avoid those naïve conclusions and get good intuition, but I want to present a more formal and objective approach.

Caveat: This still relies on assumptions (e.g.: A priori, the probability of being B is the same for “all Internet users”); but at least they can be spelled out, discussed and justified.



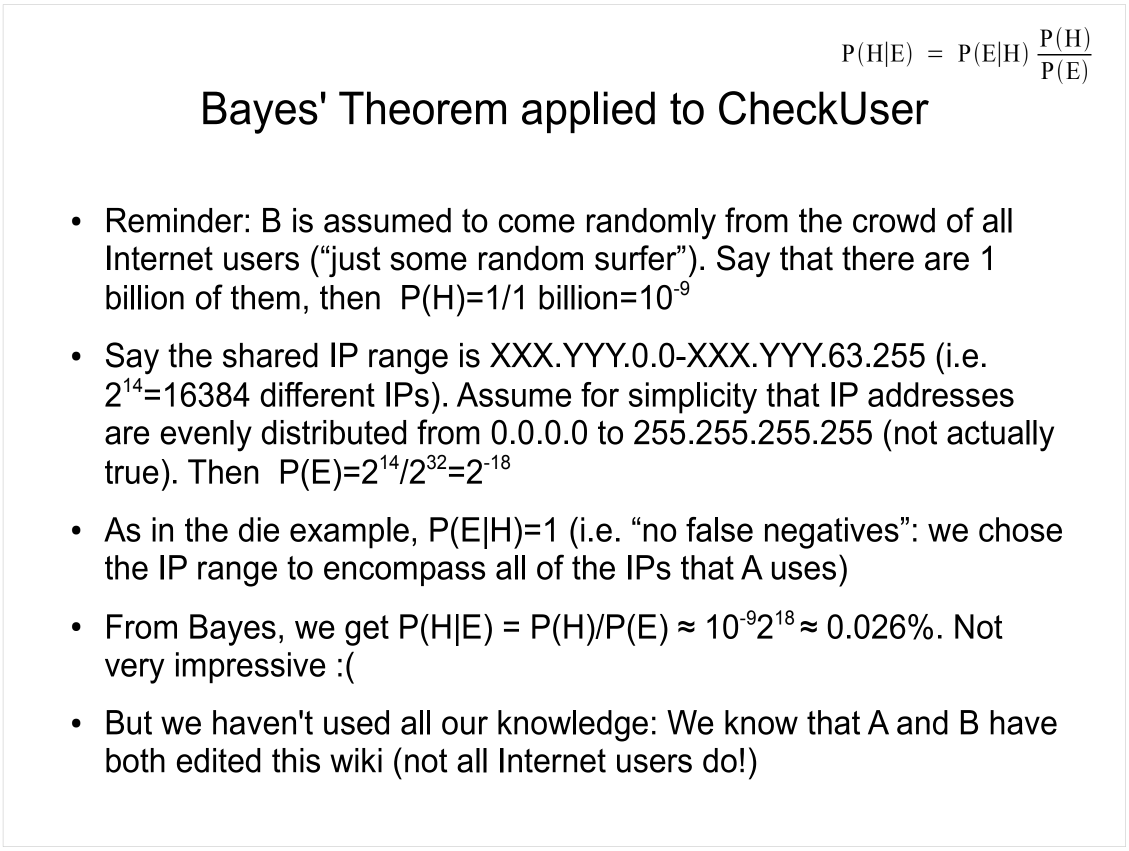
Bayes' Theorem

$$P(H|E) = P(E|H)\frac{P(H)}{P(E)}$$

- Simple example:
  - B generated by throwing a fair die (B = ⚀,⚁,⚂,⚃,⚄,⚅)
  - Hypothesis H: “B is a ⚄” (i.e. A=⚄ in above notation)
  - Evidence E: “B is odd”
  - P(E|H) = 1 (because 5 is always an odd number), P(E) = ½ (half of all numbers are odd), P(H) = 1/6 (die is fair)
  - With Bayes: P(H|E) = 1/3
- Very frequently applied in forensic statistics (DNA evidence, etc.)

So, this is the most difficult formula in this talk, promise!

http://en.wikipedia.org/wiki/Bayes%27_theorem



$$P(H|E) = P(E|H)\,\frac{P(H)}{P(E)}$$

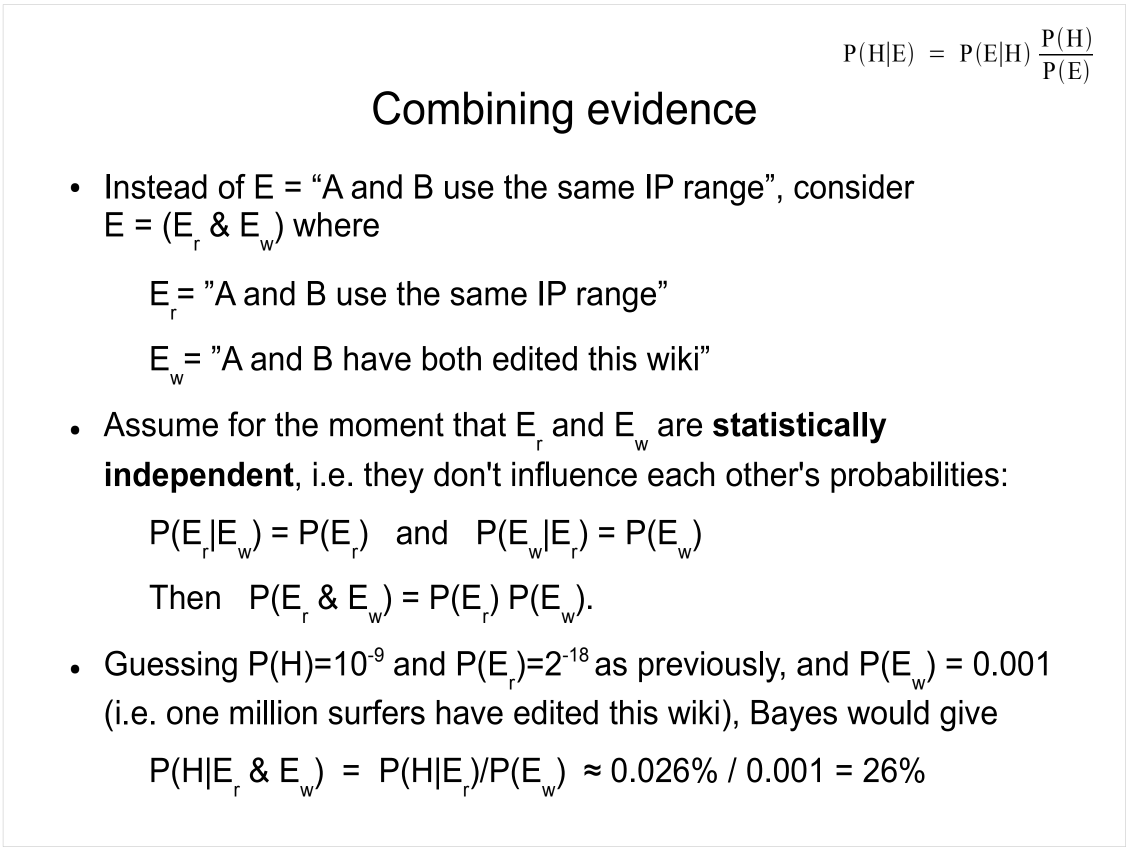
Bayes' Theorem applied to CheckUser

- Reminder: B is assumed to come randomly from the crowd of all Internet users (“just some random surfer”). Say that there are 1 billion of them, then P(H)=1/1 billion=10^{-9}
- Say the shared IP range is XXX.YYY.0.0-XXX.YYY.63.255 (i.e. 2^{14}=16384 different IPs). Assume for simplicity that IP addresses are evenly distributed from 0.0.0.0 to 255.255.255.255 (not actually true). Then P(E)=$2^{14}/2^{32}=2^{-18}$
- As in the die example, P(E|H)=1 (i.e. “no false negatives”: we chose the IP range to encompass all of the IPs that A uses)
- From Bayes, we get P(H|E) = P(H)/P(E) ≈ $10^{-9}2^{18}$ ≈ 0.026%. Not very impressive :(
- But we haven't used all our knowledge: We know that A and B have both edited this wiki (not all Internet users do!)

NB: Probability can be interpreted as knowledge, forget knowledge --> probability changes

NB: “uses IP address X” in the sense of “always uses X when accessing the Web”, not the same as “has only edited this wiki under IP address X”

To be pointlessly precise, the first and the last address in a IP range (subnet) are reserved, so the number in this example should rather be 2^{14}-2 instead of 2^{14}



$$P(H|E) = P(E|H)\frac{P(H)}{P(E)}$$

Combining evidence

- Instead of E = “A and B use the same IP range”, consider E = (E_r & E_w) where

  E_r= ”A and B use the same IP range”

  E_w= ”A and B have both edited this wiki”

- Assume for the moment that E_r and E_w are **statistically independent**, i.e. they don't influence each other's probabilities:

  $P(E_r|E_w) = P(E_r)$ and $P(E_w|E_r) = P(E_w)$

  Then $P(E_r \& E_w) = P(E_r)\,P(E_w)$.

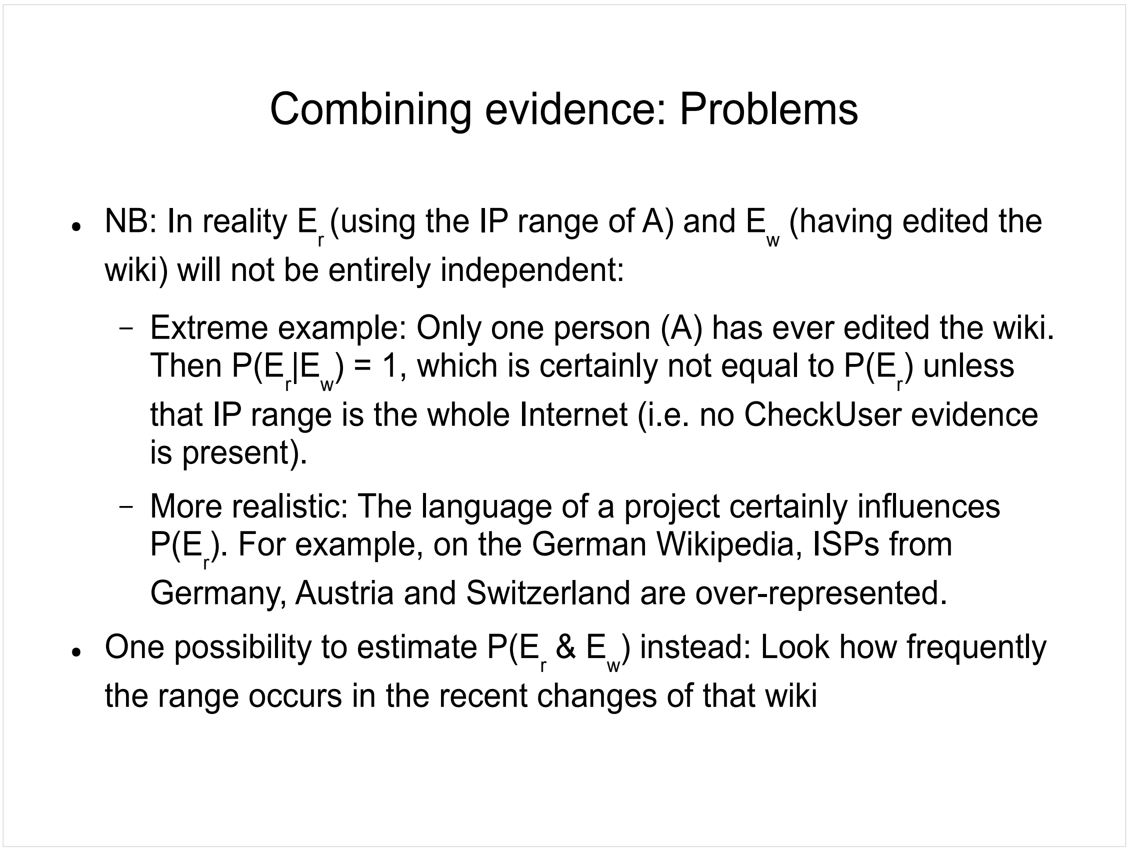
- Guessing $P(H)=10^{-9}$ and $P(E_r)=2^{-18}$ as previously, and $P(E_w) = 0.001$ (i.e. one million surfers have edited this wiki), Bayes would give

  $P(H|E_r \& E_w) = P(H|E_r)/P(E_w) \approx 0.026\% / 0.001 = 26\%$

one billion (“the whole Internet”) times 0.001 = one million

This result (26%) is much “better” than that on the previous slide, because we have used more knowledge.

Note: For a smaller wiki, $P(E_w)$ would be smaller, and Bayes' formula could give a probability greater than one! In that case, the independence assumption must have been wrong, see next slide.



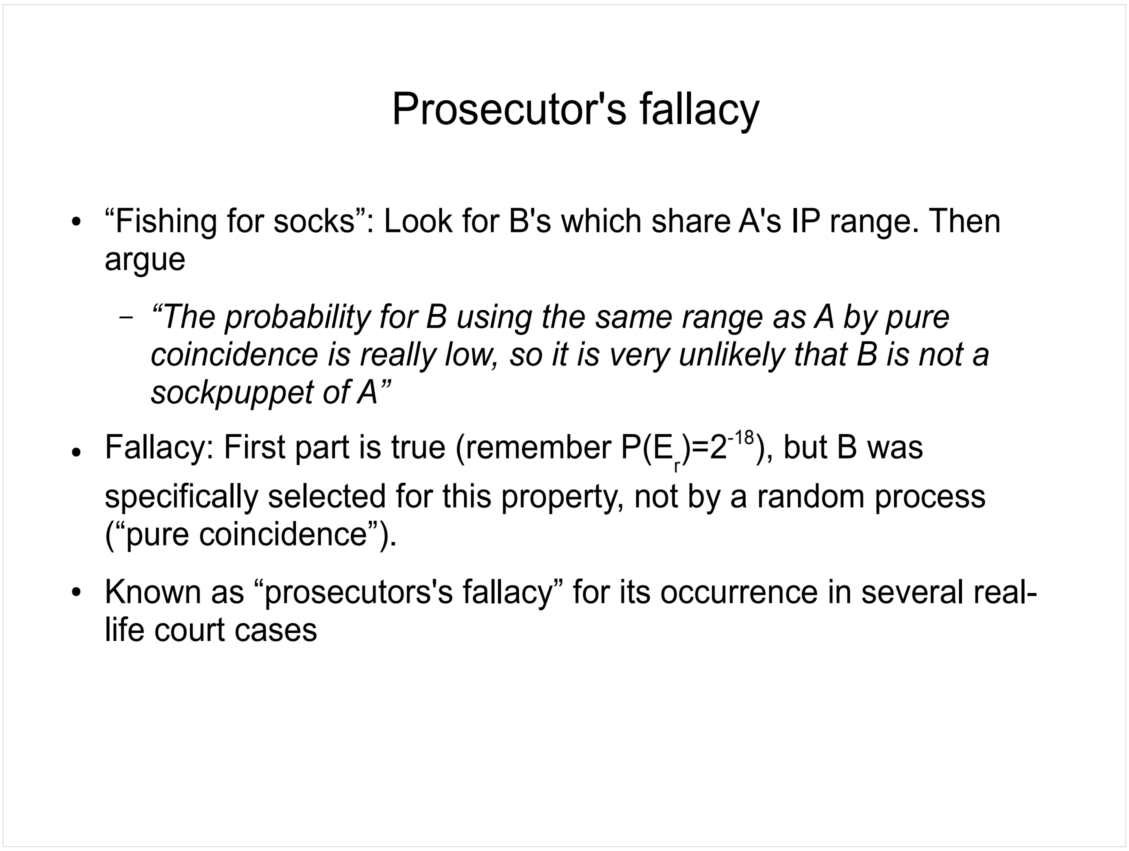
Combining evidence: Problems

- NB: In reality E_r (using the IP range of A) and E_w (having edited the wiki) will not be entirely independent:
  - Extreme example: Only one person (A) has ever edited the wiki. Then $P(E_r|E_w) = 1$, which is certainly not equal to $P(E_r)$ unless that IP range is the whole Internet (i.e. no CheckUser evidence is present).
  - More realistic: The language of a project certainly influences $P(E_r)$. For example, on the German Wikipedia, ISPs from Germany, Austria and Switzerland are over-represented.
- One possibility to estimate $P(E_r \,\&\, E_w)$ instead: Look how frequently the range occurs in the recent changes of that wiki

Can use recent changes list restricted to not logged in editors if one doesn't want to do a range CU check

Warning: In this approach, still a lot is subjective and assumptions unproven. For example, why start with “all Internet users” - why not “all German speakers” or “all humans”, or “all humans plus Martians dialing into Earth Internet”?

For more on the role of such assumptions, see http://en.wikipedia.org/wiki/Bayesian_inference#Evidence_and_changing_beliefs



Prosecutor's fallacy

- “Fishing for socks”: Look for B's which share A's IP range. Then argue
  - *“The probability for B using the same range as A by pure coincidence is really low, so it is very unlikely that B is not a sockpuppet of A”*
- Fallacy: First part is true (remember $P(E_r)=2^{-18}$), but B was specifically selected for this property, not by a random process (“pure coincidence”).
- Known as “prosecutors's fallacy” for its occurrence in several real-life court cases

Argument confuses P(H|E) with 1 - P(E|not H)

Combining with non-CU evidence, defendant's fallacy

- Recall that in our numerical example, P(H|E) was still small (nowhere near 1), even when combining IP range and being an editor
- Other evidence from CU: User agents and temporal patterns (e.g. A uses an IP at 12:07 and 12:20 pm, and B the same IP at 12:12 pm). Sometimes sufficient to conclude sockpuppetry, but:
- Usually, the CU output has to be complemented by other evidence to reach a sound conclusion.
- “Defendant's fallacy”: B argues “Tens of thousands of other people use this IP range besides me and A. So P(A=B) < 0.01%.” - Ignores that other evidence may be present.

Similar interests

- Just a few personal interests and cultural preferences can suffice to identify an individual (e.g. Narayanan,Shmatikov: *How to Break Anonymity of the Netflix Prize Dataset,* 2007)
- Frequent argument in sockpuppet cases on Wikipedia: “Both accounts edit articles about (special topic X) and (unrelated special topic Y)”
- Afaik no systematic analysis yet. But there is a tool which, for two users, displays articles that both have edited

A few movie ratings outside the mainstream (Top 100) uniquely characterize a Netflix/IMDB member
http://arxiv.org/abs/cs/0610105

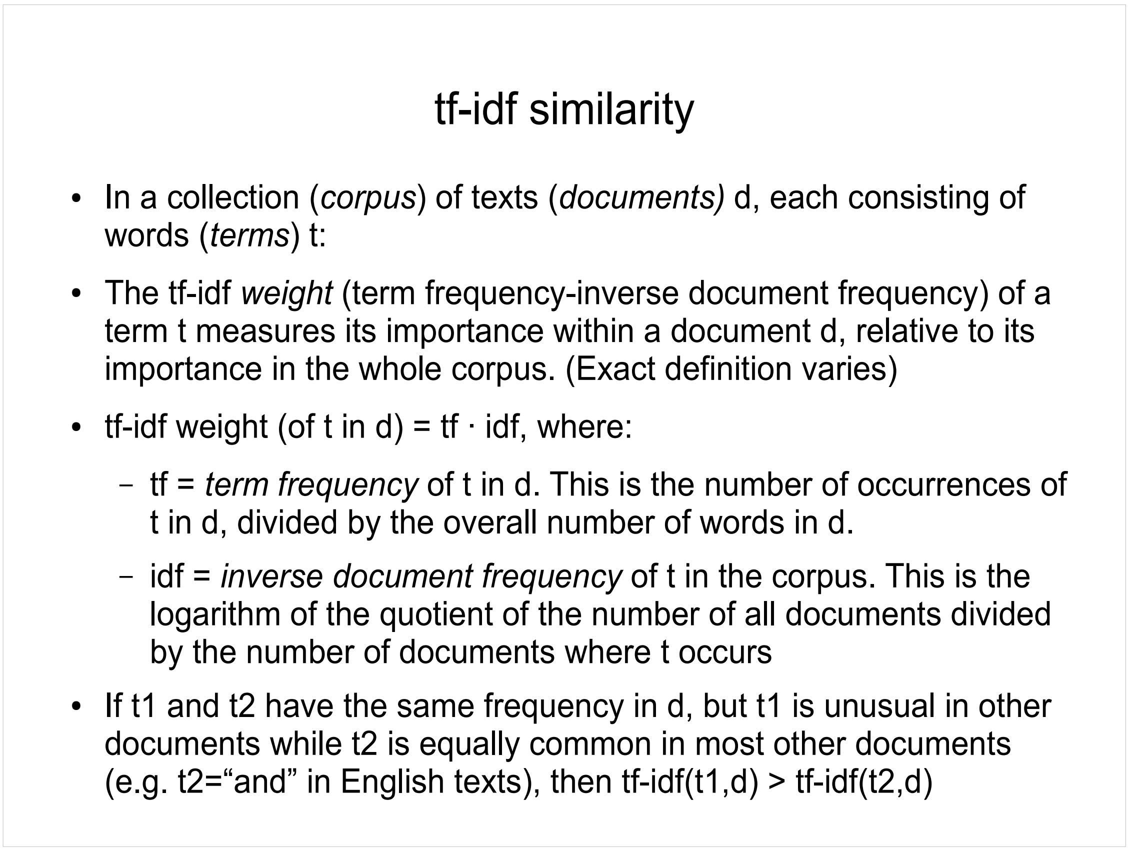
Screenshot from http://toolserver.org/~cyroxx/familiar/familiar.php (tool by user CyRoXX on dewiki, currently only available on the German Wikipedia)

Style analysis

- Users try to find significant similarities in the language used by suspected sockpuppets, such as repeated unusual typos, peculiar abbreviations, punctuation habits etc.

Stylometry, forensic linguistics

- History: Attempts to determine authorship of Shakespeare's works, the Federalist papers, the Unabomber manifesto...
- Underlying assumption: While people vary their writing style according to occasion, genre, mood etc., there exist persistent habits and traits which distinguish individual writers.
- Which properties can be regarded as persistent is often controversial, but many successes
- How does it work? Example: “tf-idf similarity”

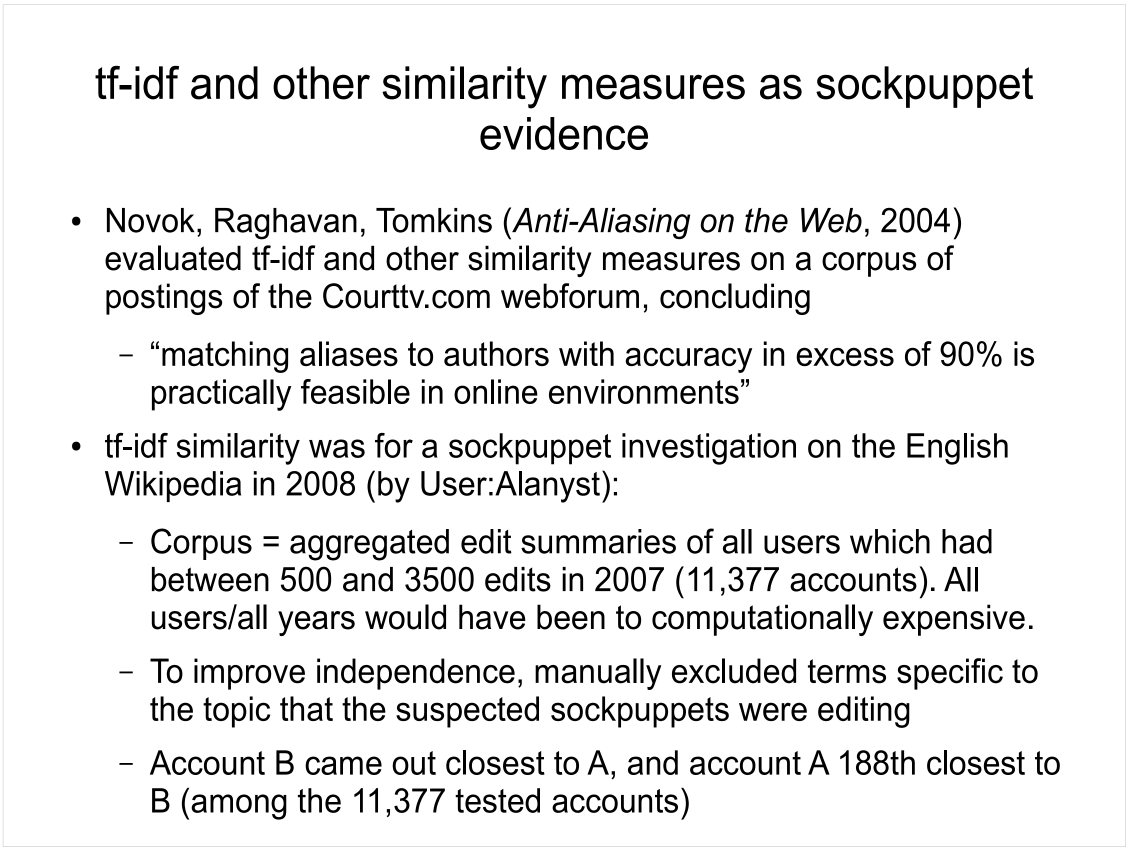


tf-idf similarity

- In a collection (*corpus*) of texts (*documents)* d, each consisting of words (*terms*) t:
- The tf-idf *weight* (term frequency-inverse document frequency) of a term t measures its importance within a document d, relative to its importance in the whole corpus. (Exact definition varies)
- tf-idf weight (of t in d) = tf · idf, where:
  - tf = *term frequency* of t in d. This is the number of occurrences of t in d, divided by the overall number of words in d.
  - idf = *inverse document frequency* of t in the corpus. This is the logarithm of the quotient of the number of all documents divided by the number of documents where t occurs
- If t1 and t2 have the same frequency in d, but t1 is unusual in other documents while t2 is equally common in most other documents (e.g. t2="and" in English texts), then tf-idf(t1,d) > tf-idf(t2,d)

tf-idf similarity

- Listing the tf-idf weights of all terms t for one d gives a vector. Angle between two of these vectors is a measure of similarity between the two documents, regarding word usage.
- Now combine the text contributions (or the edit summaries) of an user account into a document d, and take the contributions of all accounts on the wiki as the corpus. The tf-idf vector of d says something about the vocabulary preferences of that account. Accounts with a higher tf-idf similarity are more likely to be sockpuppets of the same person.



tf-idf and other similarity measures as sockpuppet evidence

- Novok, Raghavan, Tomkins (*Anti-Aliasing on the Web*, 2004) evaluated tf-idf and other similarity measures on a corpus of postings of the Courttv.com webforum, concluding
  - “matching aliases to authors with accuracy in excess of 90% is practically feasible in online environments”
- tf-idf similarity was for a sockpuppet investigation on the English Wikipedia in 2008 (by User:Alanyst):
  - Corpus = aggregated edit summaries of all users which had between 500 and 3500 edits in 2007 (11,377 accounts). All users/all years would have been to computationally expensive.
  - To improve independence, manually excluded terms specific to the topic that the suspected sockpuppets were editing
  - Account B came out closest to A, and account A 188th closest to B (among the 11,377 tested accounts)

Actually, Novok et al. found that the Kullback-Leibler measure to yield higher accuracy than the tf-idf measure, and they used a damping factor to improve results.

“Improve independence”: One would also like to use similar interest (cf. slide 22) as evidence, but users editing the same topics are expected to use terminology specific to that topic (cf.Novok p.37-38), and maybe even pick up word usage from each other, which reduces the statistical independence of these two types of evidence.

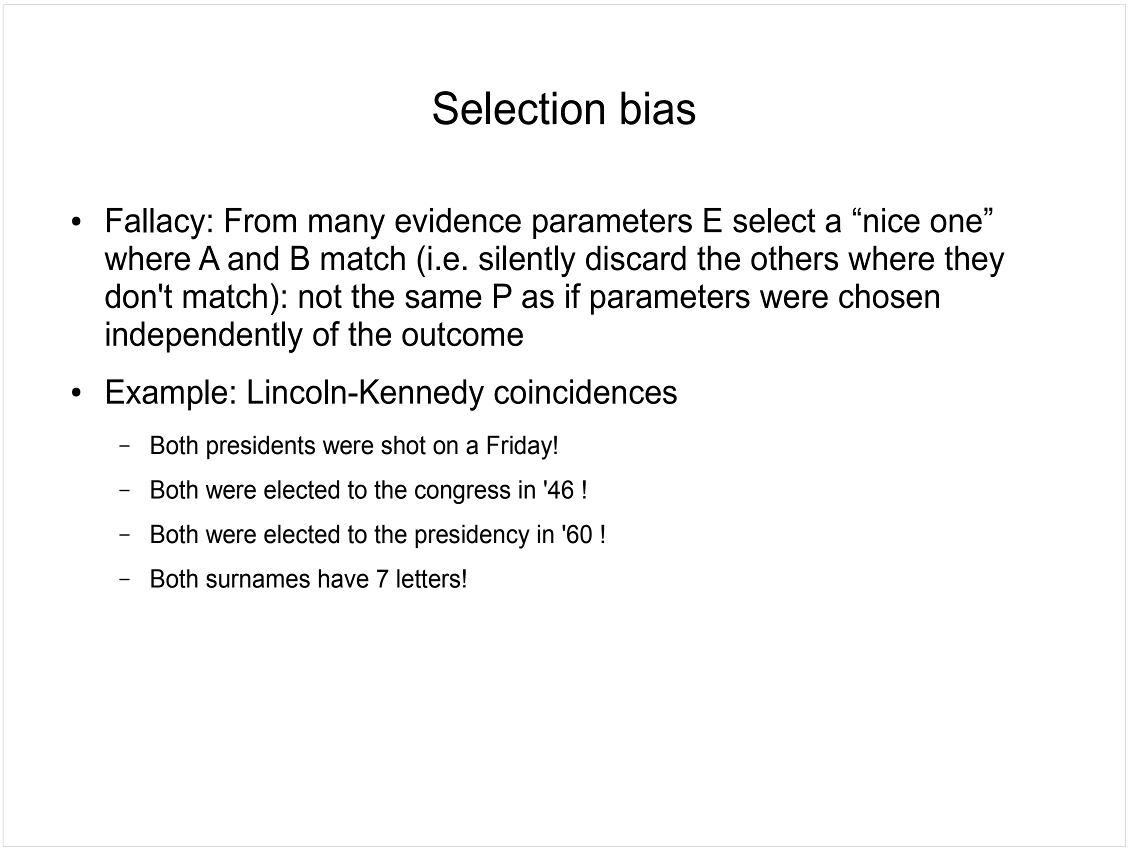
Use on many accounts simultaneously, many words each – can be computationally expensive. For the old table on enwiki, probably would be *really* expensive (cf. the WikiTrust software by the UCSC Wiki Lab). But once realized, and combined with a clustering algorithm, should be a powerful tool to uncover sockpuppets, and quite scary pricacy-wise.

The paper by Novok et al. is available at:

http://citeseerx.ist.psu.edu/viewdoc/download;?doi=10.1.1.2.3205&rep=rep1&type=pdf

Sockpuppet investigation by Alanyst:

http://en.wikipedia.org/wiki/User:Alanyst/Vector_space_research



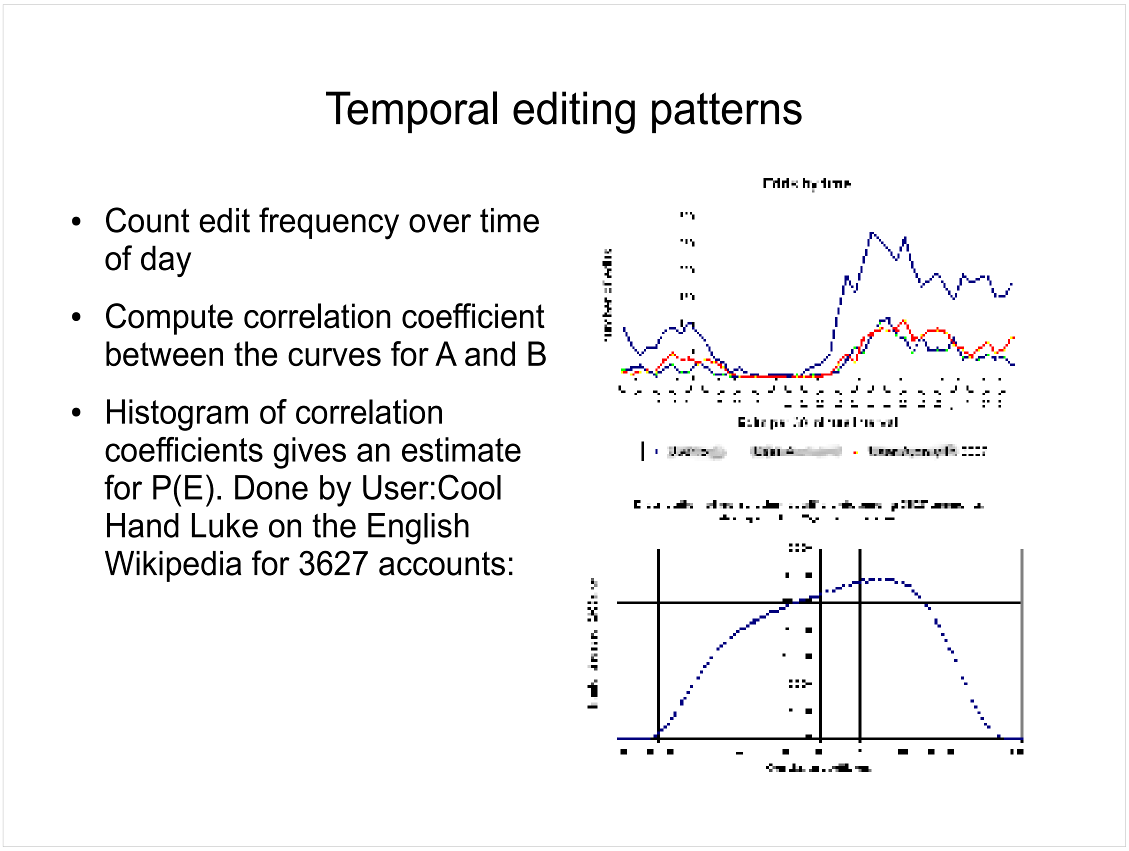
Selection bias

- Fallacy: From many evidence parameters E select a “nice one” where A and B match (i.e. silently discard the others where they don't match): not the same P as if parameters were chosen independently of the outcome
- Example: Lincoln-Kennedy coincidences
  - Both presidents were shot on a Friday!
  - Both were elected to the congress in '46 !
  - Both were elected to the presidency in '60 !
  - Both surnames have 7 letters!

Problem: Very many properties E_1,E_2,E_3 ,... are apt to be presented in such a list. Selecting only the positives (properties which coincide: E_{47} = weekday of the assassination, E_{185}= last two digits of the year of the first congress election, E_{239}= number of letters in the surname...) while silently discarding the many more negatives can create a false impression of a connection between independent things.

Analogously in sockpuppet investigations (think A=Lincoln, B=Kennedy) which examine a lot of different kinds of evidence E_1,E_2,E_3,... but discard too many negatives.

http://en.wikipedia.org/wiki/Lincoln-Kennedy_coincidences



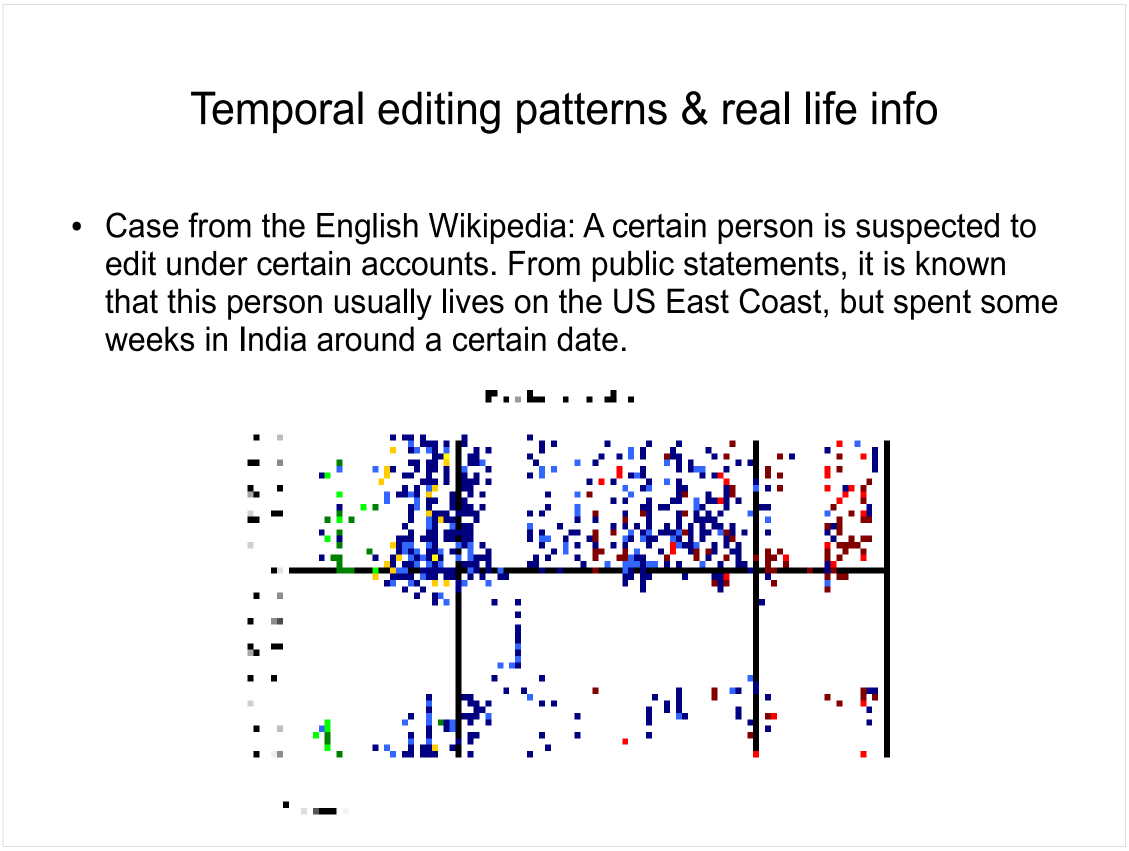
Temporal editing patterns

- Count edit frequency over time of day
- Compute correlation coefficient between the curves for A and B
- Histogram of correlation coefficients gives an estimate for P(E). Done by User:Cool Hand Luke on the English Wikipedia for 3627 accounts:

Side note: Jimmy Wales recently suggested to investigate if there was a correlation of the usage of different English dialects (US, UK, AUS) with the different times of the day

http://wikipediaweekly.org/2008/05/08/episode-48-interview-wjimmy-wales/

Diagrams adapted from
http://en.wikipedia.org/w/index.php?oldid=208039584
http://en.wikipedia.org/wiki/Image:Correlation_coefs_3627.png
Author: Cool Hand Luke, License: CC-BY 3.0

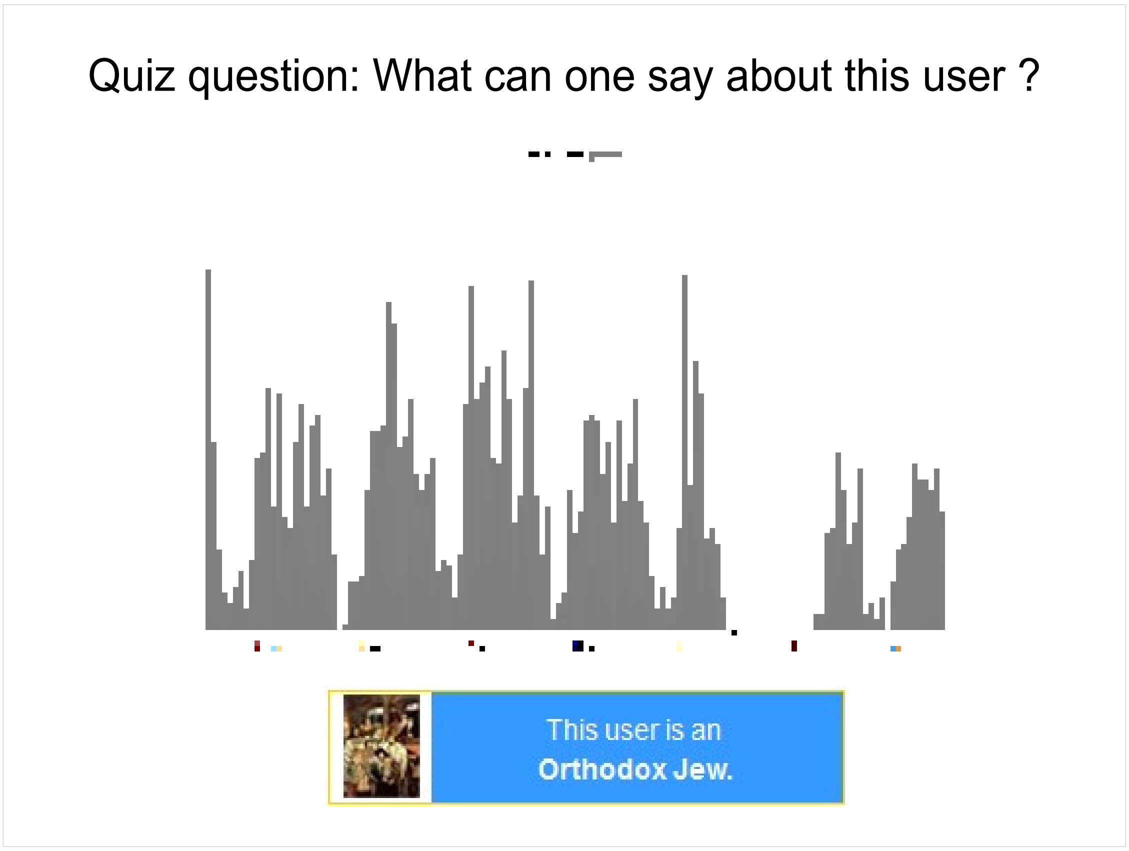


Temporal editing patterns & real life info

- Case from the English Wikipedia: A certain person is suspected to edit under certain accounts. From public statements, it is known that this person usually lives on the US East Coast, but spent some weeks in India around a certain date.

India = UTC+5:30
EST = UTC-5:00

Diagrams adapted from
http://en.wikipedia.org/w/index.php?oldid=208041005
Author: Cool Hand Luke, License: CC-BY 3.0



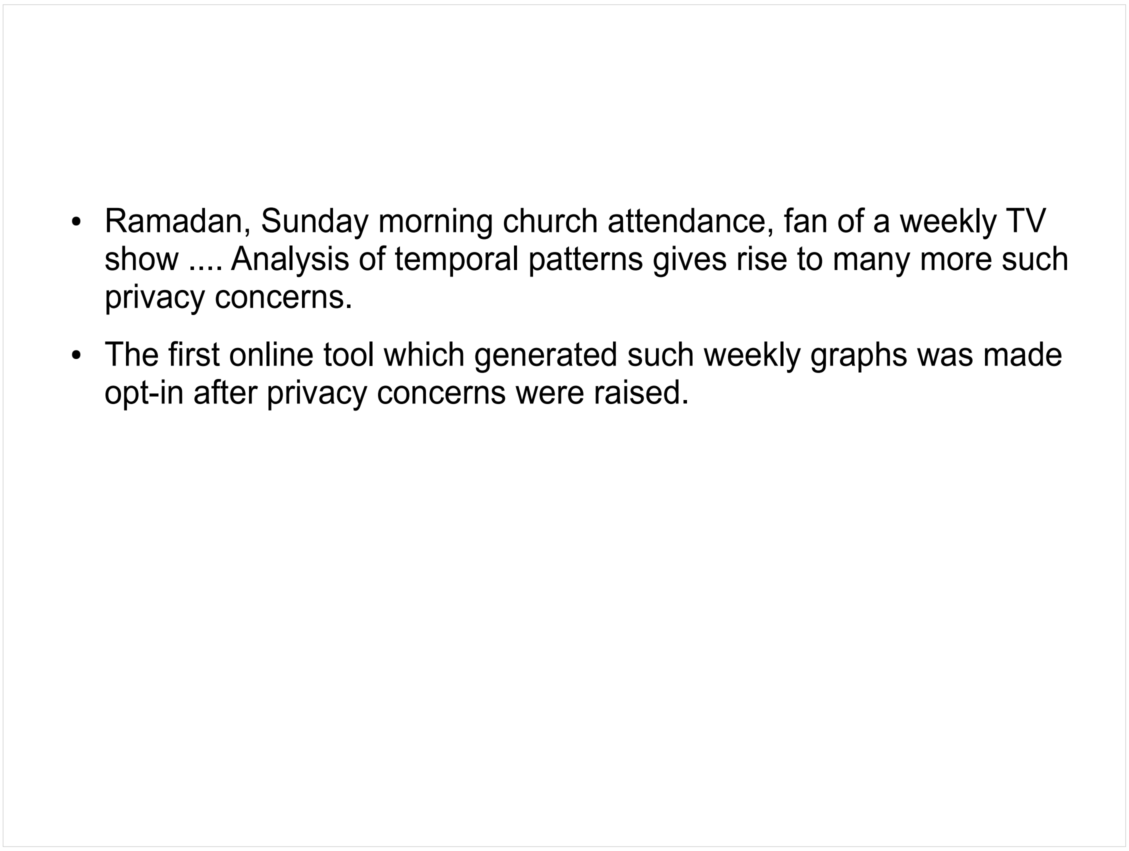
Quiz question: What can one say about this user ?

This user is an **Orthodox Jew.**

Accumulation of >3500 edits, more than three years

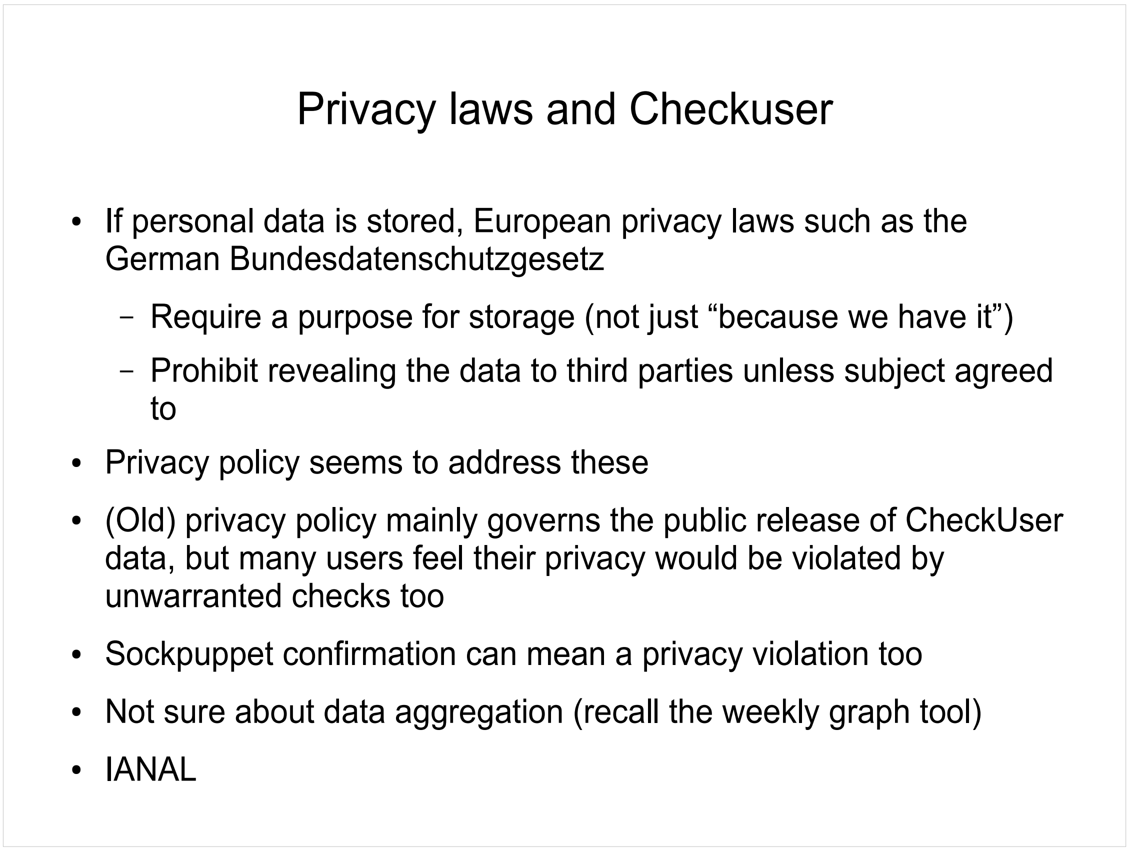
In this case, the user voluntarily disclosed his religious affiliation via a “user box” on his user page. But many other users don't want such information to be public, and it is entirely possible to write an automated program which identifies most users on a wiki who are observing the Jewish shabbat in this way.

Diagram created using Flcelloguy's Tool: http://en.wikipedia.org/wiki/Wikipedia:WPEC/FT/H



- Ramadan, Sunday morning church attendance, fan of a weekly TV show .... Analysis of temporal patterns gives rise to many more such privacy concerns.
- The first online tool which generated such weekly graphs was made opt-in after privacy concerns were raised.

A former admin of de: threatened to sue Wikimedia Deutschland (who runs the toolserver) for privacy violation because of the weekly graphs provided by “Interiot's tool”.



Privacy laws and Checkuser

- If personal data is stored, European privacy laws such as the German Bundesdatenschutzgesetz
  - Require a purpose for storage (not just “because we have it”)
  - Prohibit revealing the data to third parties unless subject agreed to
- Privacy policy seems to address these
- (Old) privacy policy mainly governs the public release of CheckUser data, but many users feel their privacy would be violated by unwarranted checks too
- Sockpuppet confirmation can mean a privacy violation too
- Not sure about data aggregation (recall the weekly graph tool)
- IANAL

Some debate on whether IP addresses are to be regarded as personal data (Google took both positions ;-)

But also: German law grants rights to the person (on whom the data is about) to request information about the stored data and even its deletion

Conclusions

- The community has developed a lot of expertise, some homegrown tools and clever techniques to generate sockpuppet evidence.
- There is still much more potential for more formalized and more automated analysis, applying research from several fields.
- Wikipedia contributors don't just give their time to the project, but pay with their privacy, too.