Privacy, openness, trust and transparency on Wikipedia

How the free encyclopedia project deals with sockpuppets

26C3, Berlin, 27 December 2009

HaeB

haebwiki@gmail.com

Please don't take photos during this talk.

Wikipedia's radical openness

- Founded in 2001, today the largest encyclopedia for many languages (German Wikipedia celebrates 1 million articles today)
- Growth enabled by the radical openness of wikis as a collaboration platform:
    - Anyone can edit, even without an account (and anyone can revert an edit ;)
    - Anyone can get an account.
    - No ID verification when registering (not even email address required)

Wikipedia's internal reputation system

- Per Wikipedia's policies, the reliability of its content should not rely on formal qualifications of the contributors (their external reputation). Instead, it should be ensured by citing reputed *sources* and collaborative error correction.
- However, any collaborative online group needs measures of trust and reputation to keep it functioning (as argued in Clay Shirky's highly recommended 2003 text “A Group Is Its Own Worst Enemy“, also one of the first to examine Wikipedia's model).
- On Wikipedia, users get formally assigned trust (e.g. adminship) or have it removed (blocking) based on their actions on the site – they determine their internal reputation.
- Shirky also observed: *“in all successful online communities that I've looked at, a core group arises that cares about and gardens effectively. Gardens the environment, to keep it growing, to keep it healthy.“*
- Has Wikipedia's “core group” become too close-minded and defensive? An eternal debate on Wikipedia. I am leaving this to the panel discussion (Wednesday 11:30, Saal 1). Here, focus is on structures, methods, and tools of gardening.

What are sockpuppets?

- Multiple accounts controlled by the same user
- Many legitimate uses for multiple accounts
- “Sockpuppet” often implies deceptive intention, think ventriloquist

What is the problem with sockpuppets?

- “**Sybil attacks**”:
  - Ballot stuffing in votes (a few non-content WP decision processes rely on voting, such as request for adminships and elections to the Board of Trustees of the Wikimedia Foundation).
  - More generally, artificial majorities in content disputes, especially circumventing the “three-revert-rule” on English Wikipedia.
- “**Dr Jekyll/Mr Hyde**”: Carry out evil or controversial actions with a sockpuppet, main account remains in good standing:
  - Trolling (actions intended to provoke adversive r eactions and disrupt the community)
  - Strawman accounts (putting the adversarial position in bad light)
- **Ban evasion**

What is the problem with sockpuppets?

(cont'd)

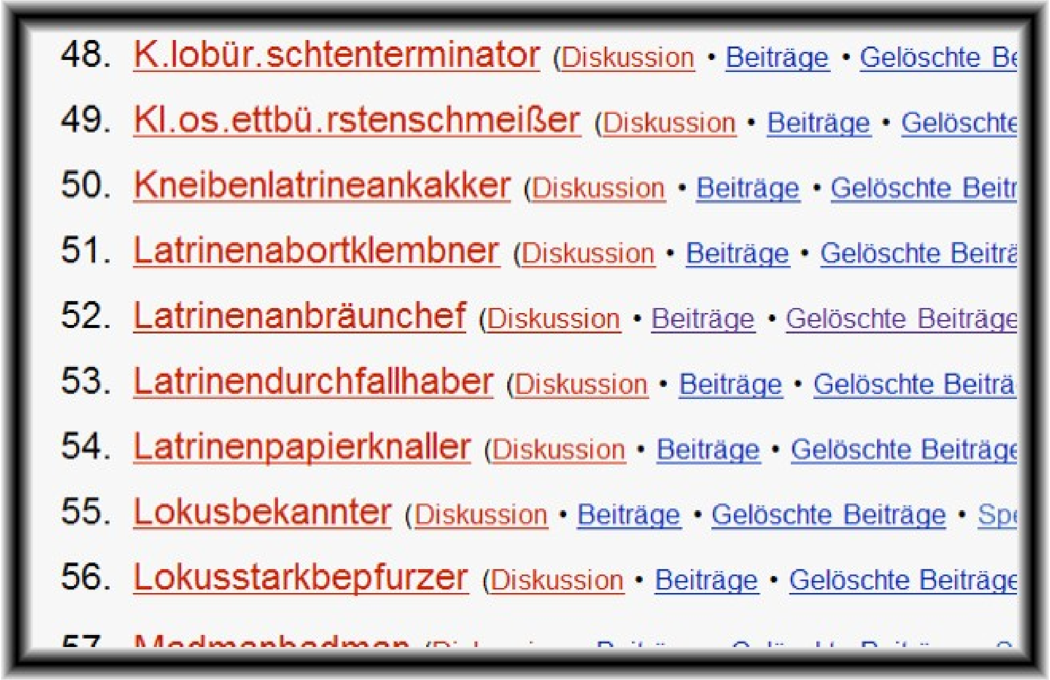
- **Newbies** get treated badly because of the possibility that they might be a banned user returning as a sockpuppet:
    - Friedman and Resnick (*The Social Cost of Cheap Pseudonyms*, Journal of Economics and Management Strategy 2001): Proved in a game-theoretic model that the possibility of creating 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.

Ban evasion

- Autoblock - a mechanism in MediaWiki preventing the simplest form of ban evasion: For an edit attempt with a blocked account, the used IP is blocked too (for 24h), and also other accounts using that IP.
- On a dialup connection with dynamically assigned IP, this is easily defeated by reconnecting and obtaining another IP in the same range.
- Autoblock can already lead to collateral damage (especially on some ISPs which rotate dynamic IPs very quickly, like AOL).
- For more sophisticated detection of ban evasion, need manual inspection of the IPs of the suspected sockpuppet before deciding about the block.

Range blocks

- For severe cases of mass vandalism, blocking the whole range is an option.
- Problem: Potentially huge collateral damage. Can be estimated by looking at previous “good” edits from that range. If a hard block is considered, logged-in edits are of interest too - needs a range CU check, problematic



48. K.lobür.schtenterminator (Diskussion • Beiträge • Gelöschte Be
49. Kl.os.ettbü.rstenschmeißer (Diskussion • Beiträge • Gelöschte
50. Kneibenlatrineankakker (Diskussion • Beiträge • Gelöschte Beitr
51. Latrinenabortklembner (Diskussion • Beiträge • Gelöschte Beiträ
52. Latrinenanbräunchef (Diskussion • Beiträge • Gelöschte Beiträge
53. Latrinendurchfallhaber (Diskussion • Beiträge • Gelöschte Beiträ
54. Latrinenpapierknaller (Diskussion • Beiträge • Gelöschte Beiträge
55. Lokusbekannter (Diskussion • Beiträge • Gelöschte Beiträge • Sp
56. Lokusstarkbepfurzer (Diskussion • Beiträge • Gelöschte Beiträge

Excerpt from a list of many thousand sock puppets created on de:WP by a vandal with a toilet fixation (since at least 2006, still active – one of the few cases where range blocks are used regularly)

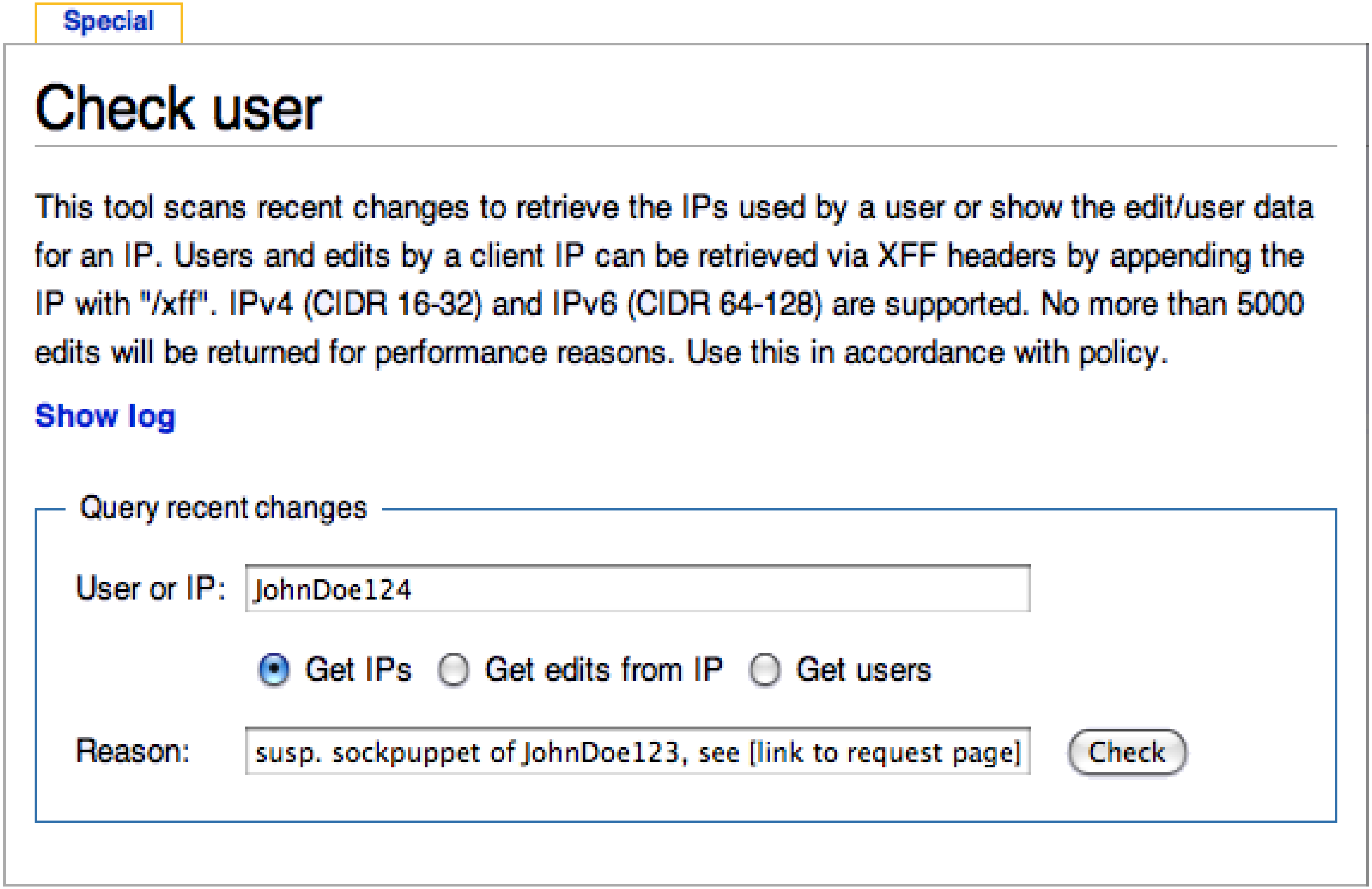
The Checkuser tool in MediaWiki

- Allows a few trusted users on a wiki to manually inspect IP addresses from which edits are made from (not: reader's IPs)
- On many web sites and web forums, administrators can routinely see participant's IPs, email adresses etc.
- Wikimedia Foundation's privacy policy sets higher expectations.
- Access to the CheckUser tool is only granted to a few trusted users on each wiki. Must be over 18 years and identify themselves to the Wikimedia Foundation
- Each access is logged (but log is only visible to other CheckUsers due to privacy concerns)
- Ombudsman commission for investigation of alleged violations of the privacy policy

Local CheckUser policies

- Use of the CheckUser tool differs on various Wikipedias due to local policies and circumstances.
- Oct 2006-Dec 2007: ca. 33.000 checks on en:WP, ca 1.100 on de:WP.
- As of December 2009, 3 CheckUsers on German Wikipedia, 36 on English Wikipedia.
- On German Wikipedia:
  - CheckUsers only perform inspections on request by other users, don't investigate on their own. Only done if there is already significant other evidence for sockpuppetry
  - All requests are publicly noted, usually naming the checked accounts (but not the private IP data)
  - Blocking actions based on CU results are left to other admins

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

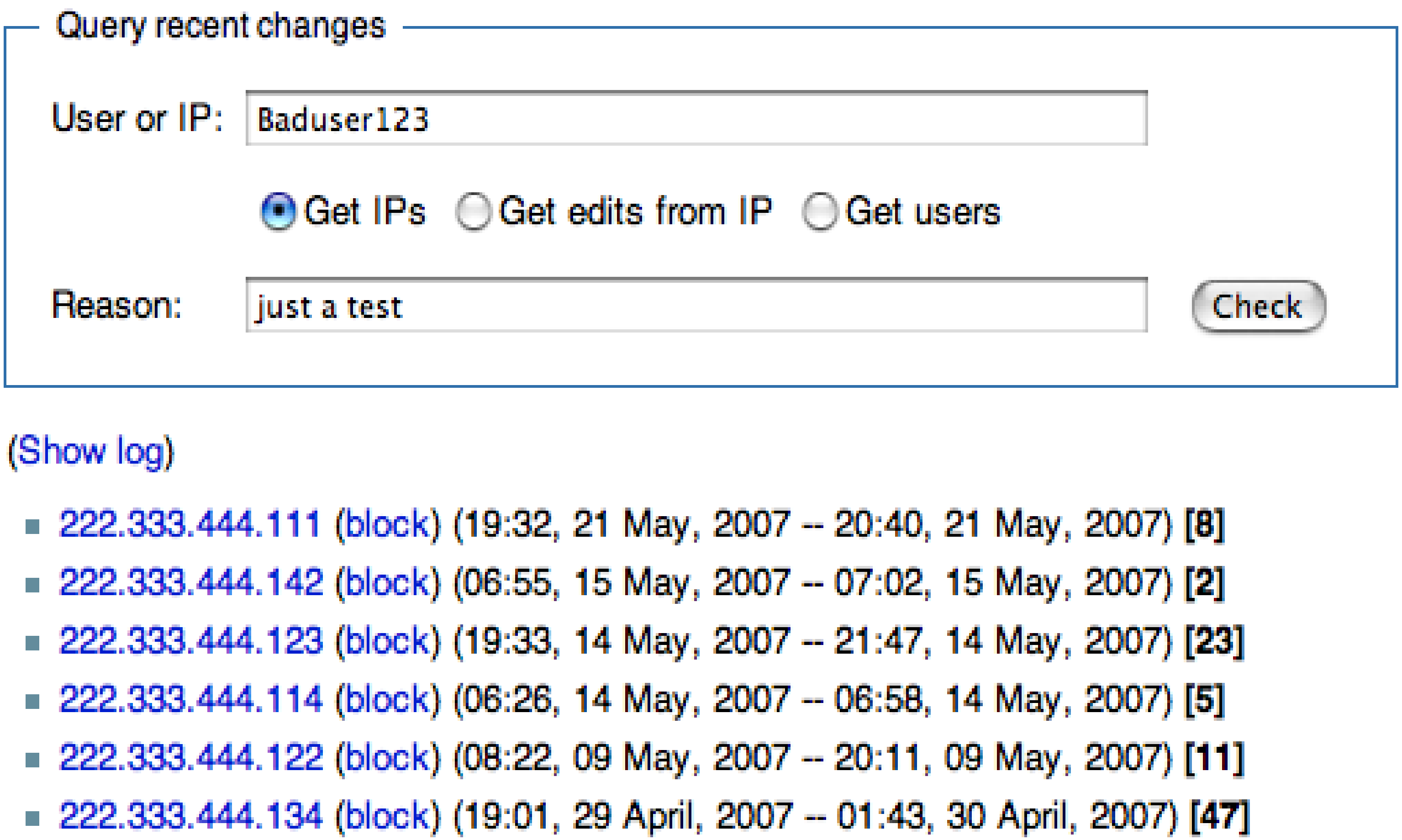
Query recent changes

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

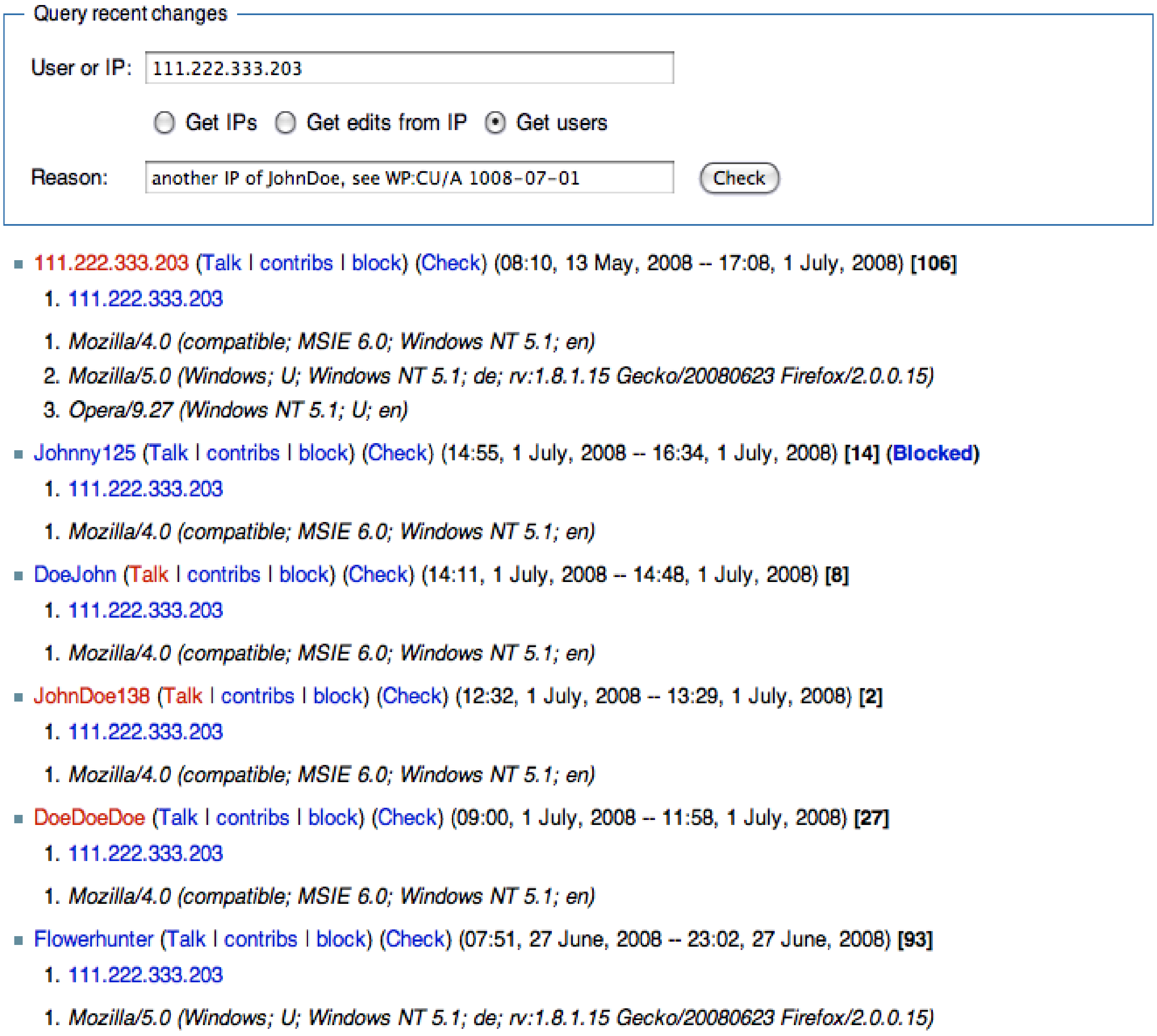
● Get IPs ○ Get edits from IP ○ Get users

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 users editing from an IP



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)

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, ...)
  - NAT: The same entry point is often used by more than one computer. (The 2006 “Illuminati” study by Casado and Freedman found this to be the case for ca. 60% of web clients, but fortunately most NAT pools are small, < 7 clients.)
  - Dynamic IPs (especially 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 if E) = ?   where
  - H is the hypothesis (“A=B”)
  - E is the evidence (both use the same IP range to access the Internet)
  - P(H if 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

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

- 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 if 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 if E) = 1/3
- Very frequently applied in forensic statistics (e.g. DNA evidence)

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

Bayes' Theorem applied to CheckUser – a toy example

- Reminder: A priori, the suspected account 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$=10^{-9}$
- Let the evidence E be that A and B share an /18 IP range YYY.0.0-XXX.YYY.63.255 (i.e. $2^{14}=16384$ different IPs). Then $P(E)=2^{14}/2^{32}=2^{-18}$ (Somewhat naively assuming that IP addresses are evenly distributed from 0.0.0.0 to 255.255.255.255.)
- As in the die example, P(E if 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 if 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!)

$$P(H \text{ if } E) = P(E \text{ if } H) \frac{P(H)}{P(E)}$$

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:

  $P(E_r \text{ if } E_w) = P(E_r)$ and $P(E_w \text{ if } E_r) = P(E_w)$

  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

  $P(H \text{ if } E_r \,\&\, E_w) = P(H \text{ if } E_r)/P(E_w) \approx 0.026\% / 0.001 = 26\%$

Combining evidence: Problems

- NB: In reality E_r (using the same IP range as 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 $P(E_r \text{ if } E_w) = 1$, 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 – a surfer who uses their ranges is more likely to edit that wiki.
- 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
- Cf. Anne Roth's talk at 24C3: Police (BKA) googled two terms from a letter claiming responsibility for an arson attack (“Gentrification” and “Prekarisierung”), found a sociologist who had used them in his writing, and arrested him later.

Combining with non-CU evidence, defendant's fallacy

- Recall that in our numerical example, P(“A = B” if E) was still small (nowhere near 1), even when combining IP range and being an editor as evidence E
- Other evidence from CheckUser results: 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. Wikipedia's radical transparency means that a lot of such evidence is available from user contribution list, see next slides.
- “**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.

Selection bias

- Fallacy: From many evidence parameters E select “nice ones” 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 !

  Etc. … proving that JFK must have been ~~a sockpuppet of~~ somehow a reincarnation of Abraham Lincoln !!11!!

Style analysis

- Users frequently try to find significant similarities in the language used by suspected sockpuppets, such as repeated unusual typos, peculiar abbreviations, punctuation habits etc.
- Simple examples from actual CU cases on de:WP:
  - “Users A and B both often leave a blank before a comma”
  - “Users A and B both sometimes end messages to their adversaries with *Und Tschüss!*”
- More sophisticated analysis, as known from the academic field of stylometry, is still rare though.

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 the field has 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. (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 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

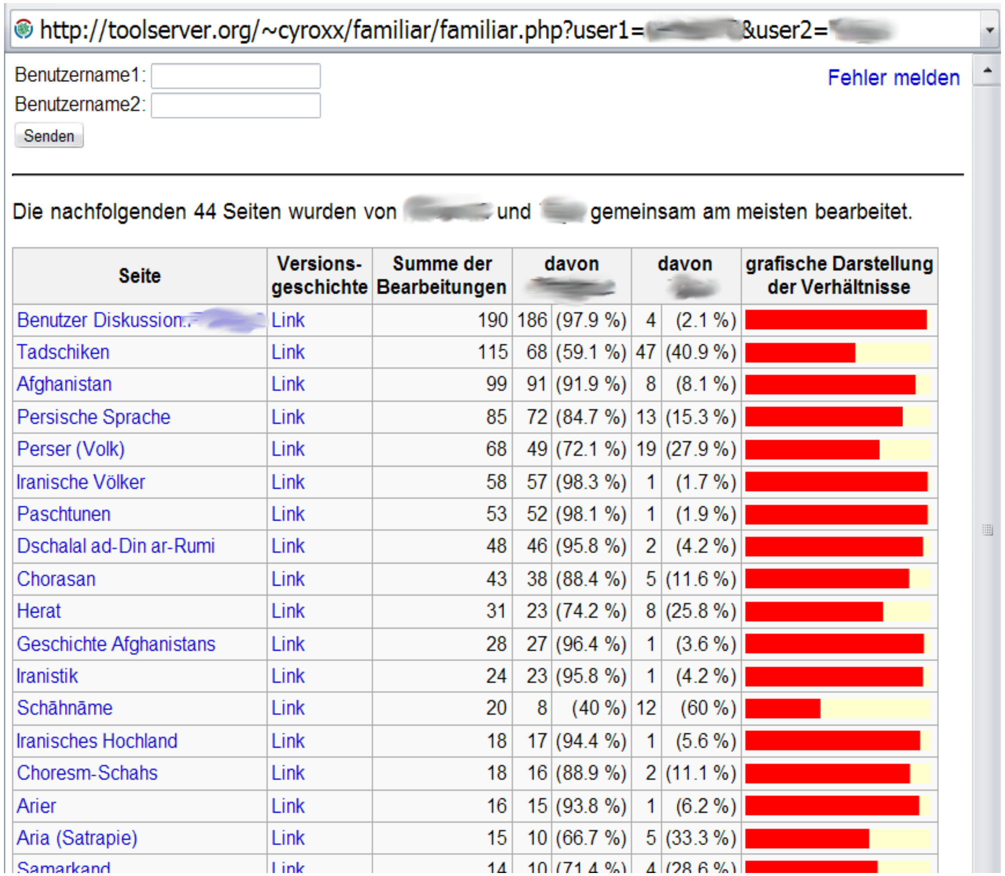
- Listing the tf-idf weights of all terms t for one d gives a vector. Angle between two of these vectors is a measure of similarity between the two documents, regarding word usage.
- Now combine the text contributions (or the edit summaries) of an user account into a document d, and take the contributions 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 too 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)

Similar interests

- Just a few personal interests and cultural preferences can suffice to identify an individual (cf. 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)”



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 %)	🟥

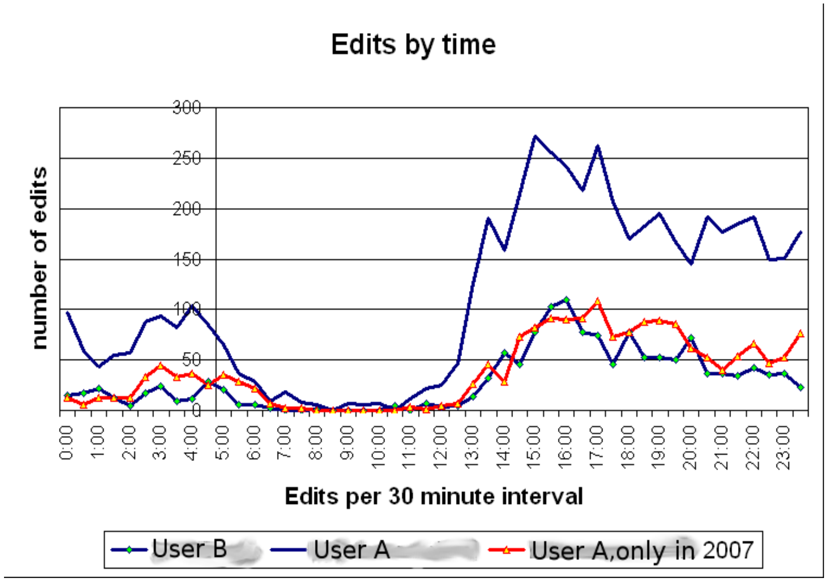
A tool which, for two users, displays articles that both have edited

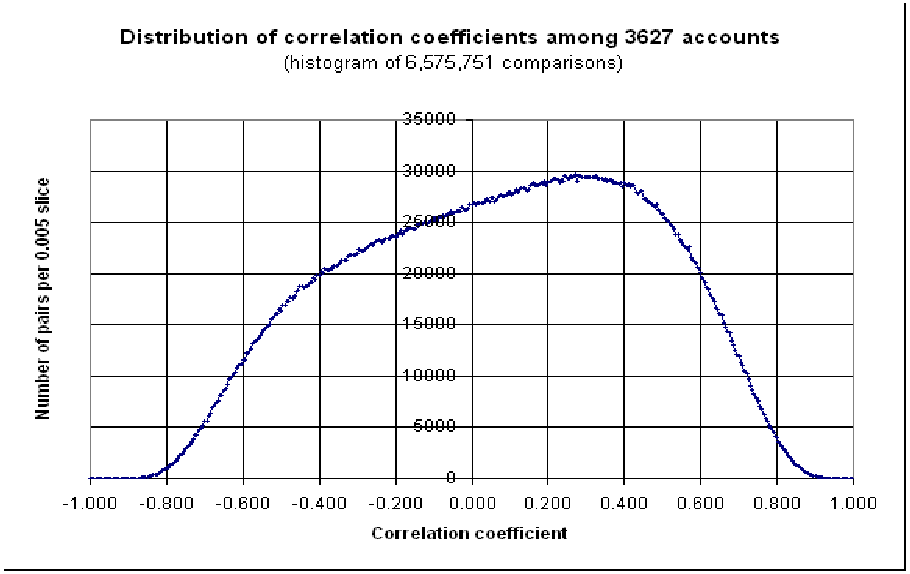
Edit collision analysis

- Assumption: A not too sophisticated sockpuppeteer will not edit with more than one account simultaneously (or say, within the same minute)
- If not purposefully avoided, this can be useful for accounts with many edits: Check for “collisions” which are becoming more likely the more edits fall within a given time span

Temporal editing patterns

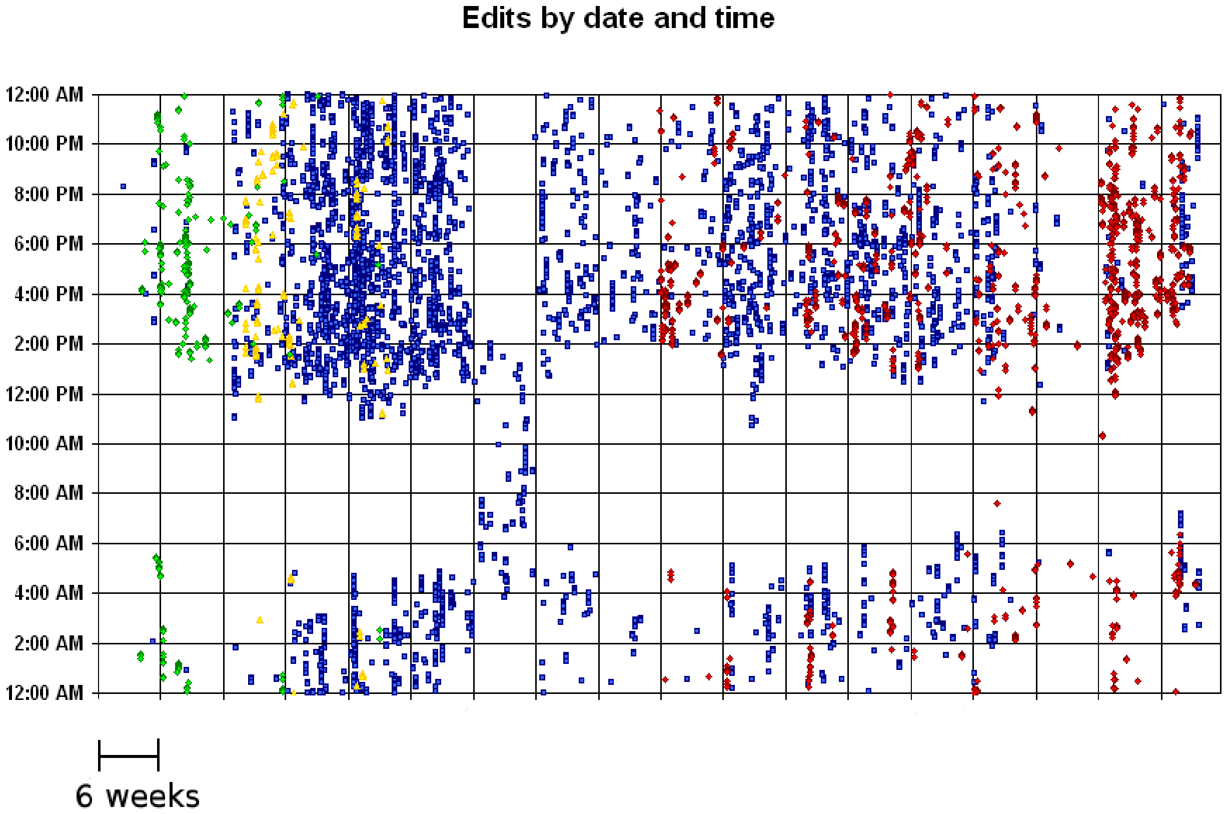
- Count edit frequency over time of day
- Compute correlation coefficient between the curves for A and B
- Evidence E = “correlation coefficient is at least as large as that of A and B”. Histogram of correlation coefficients gives an estimate for P(E). Calculated on the English Wikipedia in 2008 for 3627 accounts:



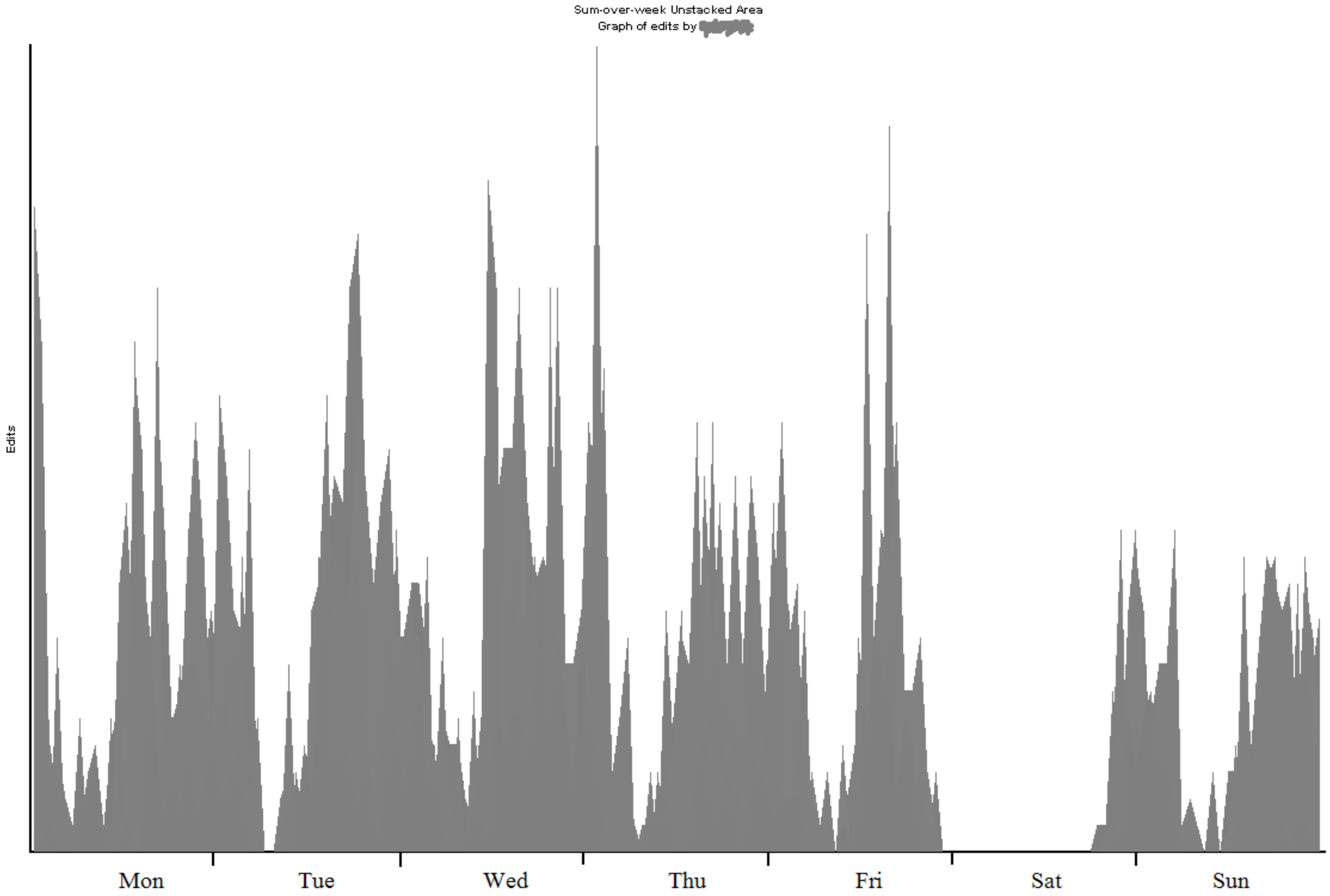


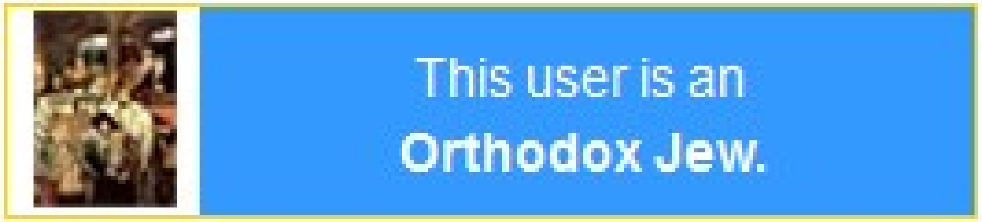
Temporal editing patterns & real life info

- 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.**

Require real names?

- Perennial proposal on Wikipedia: Abolish pseudonymous editing, require real names
- Two forks implemented this:
  - Wikiweise (started in 2005 by a former administrator and well-known deletionist of the German Wikipedia, concerned about sockpuppets, lack of quality, and excessive coverage of non-notable topics)
    - Disabled user contributions list due to privacy concerns
  - Citizendium (started in 2006 by Larry Sanger, former chief organizer of Wikipedia)
    - After self-registration lead to vandalism, switched to a stricter verification process (e.g. no freemail addresses).
- Both projects are struggling to attract enough participation. On Citizendium, anecdotal evidence that registration process is deterring many valuable potential contributors

Attempts to implement a formal web of trust on the German Wikipedia

- „**Vertrauensnetz**“ („web of trust“):
- Begun in 2004 as a trial
- Purpose: Not clearly defined – roughly: Make the community reputation of a user visible. Cited e.g. if that user runs for adminship
- Very simple technique using existing MediaWiki features:
    - If User A trusts User B, she creates a link [[Benutzer:A/Vertrauen]] → [[Benutzer:B/Vertrauen]]
    - Backlink function lists users which trust B.
- Became controversial, especially since the “/Vertrauen“ pages were also used for expressing *dis*trust in a user
- Seems to have decreased in popularity

Attempts to implement a formal web of trust on the German Wikipedia

- „**Bürgschaftsverfahren**“
- Begun in 2008
- Purpose: Certify an account as not being a sockpuppet (i.e. a person can have only one certified account)
- PIS := SHA256(Full name and birth date of owner)
  - PIS is published. Sockpuppets would have same PIS.
- Needs verification of full name and birth date by a trusted user (e.g. at a Wikipedia meetup). Started out with a few „Urbürgen“ whose identity is known to Wikimedia (i.e. Wikimedia as CA), other users rise in trust according to how many users have certified their PIS.
- Privacy problems: If one knows name + birth date, one can look up the user name. Also vulnerable to dictionary attacks (find out name+birth date from PIS, if name is not too rare).
- Has not been widely adapted (68 certified accounts as of December 2009, one year earlier: 59 certified accounts)

Attempts to implement a formal web of trust on the German Wikipedia

- „**Persönliche Bekanntschaften**“ („personal acquaintances“)
- Begun in 2008
- Purpose: Certify an account as „probably not a sockpuppet“
- Participants confirm to have met the owner of an account in real life (mostly at meet-ups), and promise not to confirm several accounts for one owner.
- Started out with a few trusted users whose identity is know to Wikimedia, others become trusted after three confirmations
- Soft security, but seems good enough
- Privacy friendly: No ID, name or other tangible real-life information required
- Facilitated by a user-side Javascript gadget (actived in preferences), a bot and a database on an external server
- Very successful: 686 participants (> 75% non-admins), 558 certified, >18000 confirmations as of December 2009. But: Impact still to be seen, certification is not (yet) a formal requirement for any function or activity in the community.

- Thanks for listening
- Questions?
- Wikipedia panel discussion (in German): Wednesday 11:30, Saal 1
- “Hack the Wiki” workshop

Privacy, openness, trust and transparency on Wikipedia

How the free encyclopedia project deals with sockpuppets

26C3, Berlin, 27 December 2009

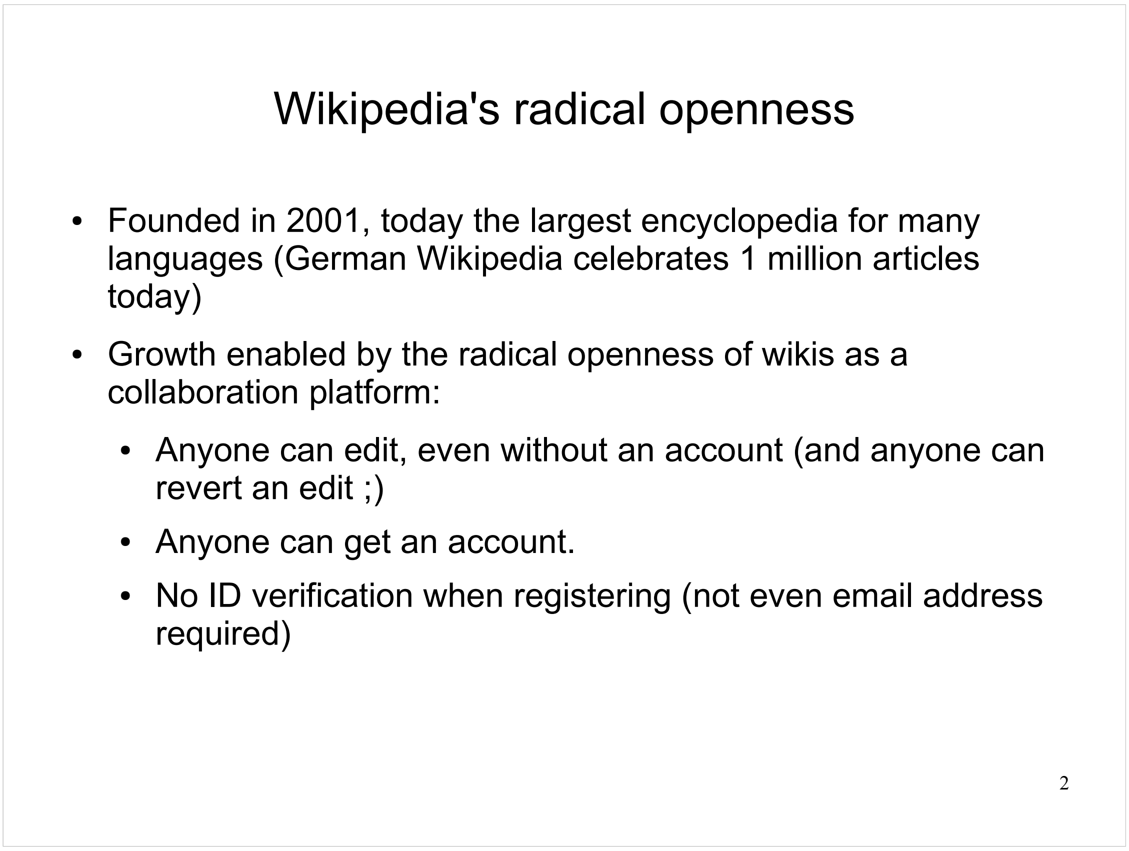
HaeB

haebwiki@gmail.com

Please don't take photos during this talk.

I apologize for some omissions in comparison to the abstract:
http://events.ccc.de/congress/2009/Fahrplan/events/3722.en.html
Among them anti-vandal bots and edit filters, the Trusted XFF project, open proxies and TORblock.

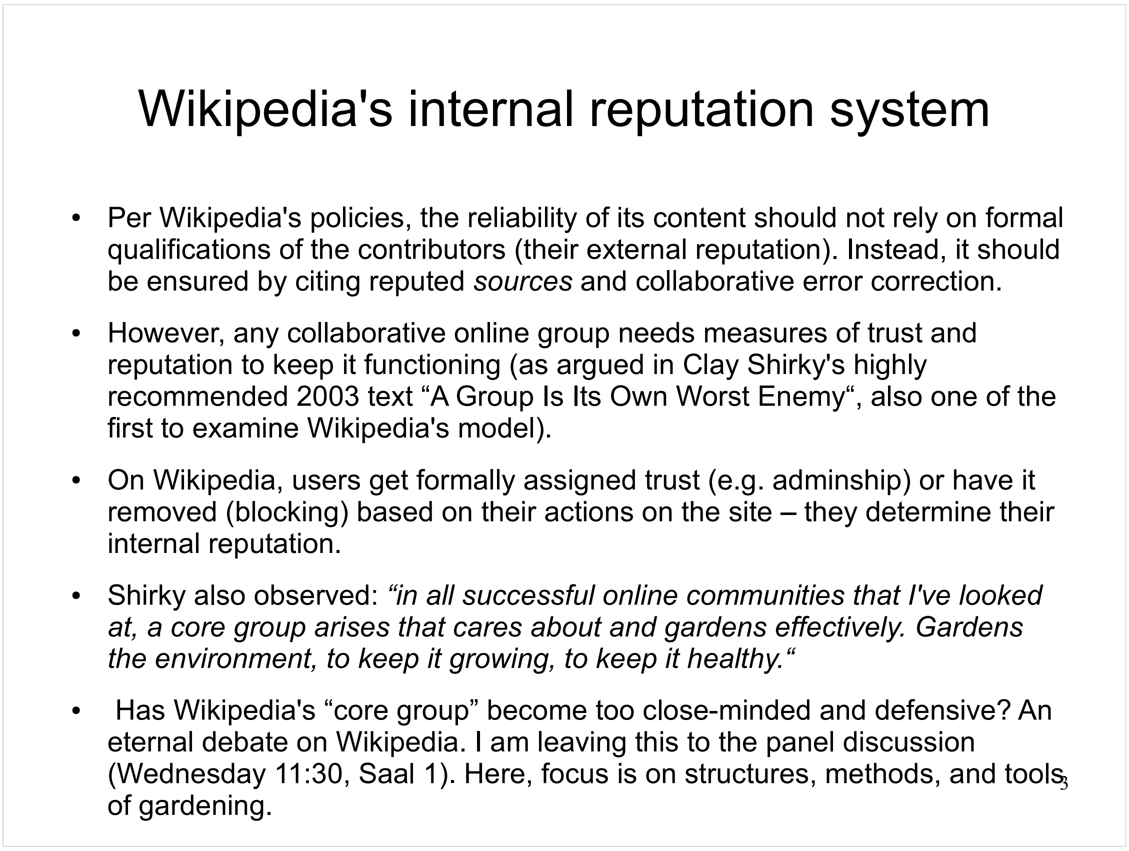
Parts of this talk correspond to my talk about a similar topic at Wikimania 2008:
http://wikimania2008.wikimedia.org/wiki/File:CheckUser_and_Editing_Patterns.pdf



Wikipedia's radical openness

- Founded in 2001, today the largest encyclopedia for many languages (German Wikipedia celebrates 1 million articles today)
- Growth enabled by the radical openness of wikis as a collaboration platform:
  - Anyone can edit, even without an account (and anyone can revert an edit ;)
  - Anyone can get an account.
  - No ID verification when registering (not even email address required)

“Anyone can edit” never meant “anyone can have the final say about the content”, though



Wikipedia's internal reputation system

- Per Wikipedia's policies, the reliability of its content should not rely on formal qualifications of the contributors (their external reputation). Instead, it should be ensured by citing reputed *sources* and collaborative error correction.
- However, any collaborative online group needs measures of trust and reputation to keep it functioning (as argued in Clay Shirky's highly recommended 2003 text “A Group Is Its Own Worst Enemy“, also one of the first to examine Wikipedia's model).
- On Wikipedia, users get formally assigned trust (e.g. adminship) or have it removed (blocking) based on their actions on the site – they determine their internal reputation.
- Shirky also observed: *“in all successful online communities that I've looked at, a core group arises that cares about and gardens effectively. Gardens the environment, to keep it growing, to keep it healthy.“*
- Has Wikipedia's “core group” become too close-minded and defensive? An eternal debate on Wikipedia. I am leaving this to the panel discussion (Wednesday 11:30, Saal 1). Here, focus is on structures, methods, and tools of gardening.

Shirky said this to explain why Wikipedia administrators are „reflexively suspicious of everyone from watching people attack Wikipedia over all the years“: *"If everyone who works at Britannica were fired, the encyclopedia would become out of date and less useful over time [...] But if everyone who really cares about defending Wikipedia didn't log in this week, it would be gone by Thursday."*

http://www.thedailybeast.com/blogs-and-stories/2009-11-28/wikipedias-attack-dog-editors/full/

What are sockpuppets?

- Multiple accounts controlled by the same user
- Many legitimate uses for multiple accounts
- “Sockpuppet” often implies deceptive intention, think ventriloquist

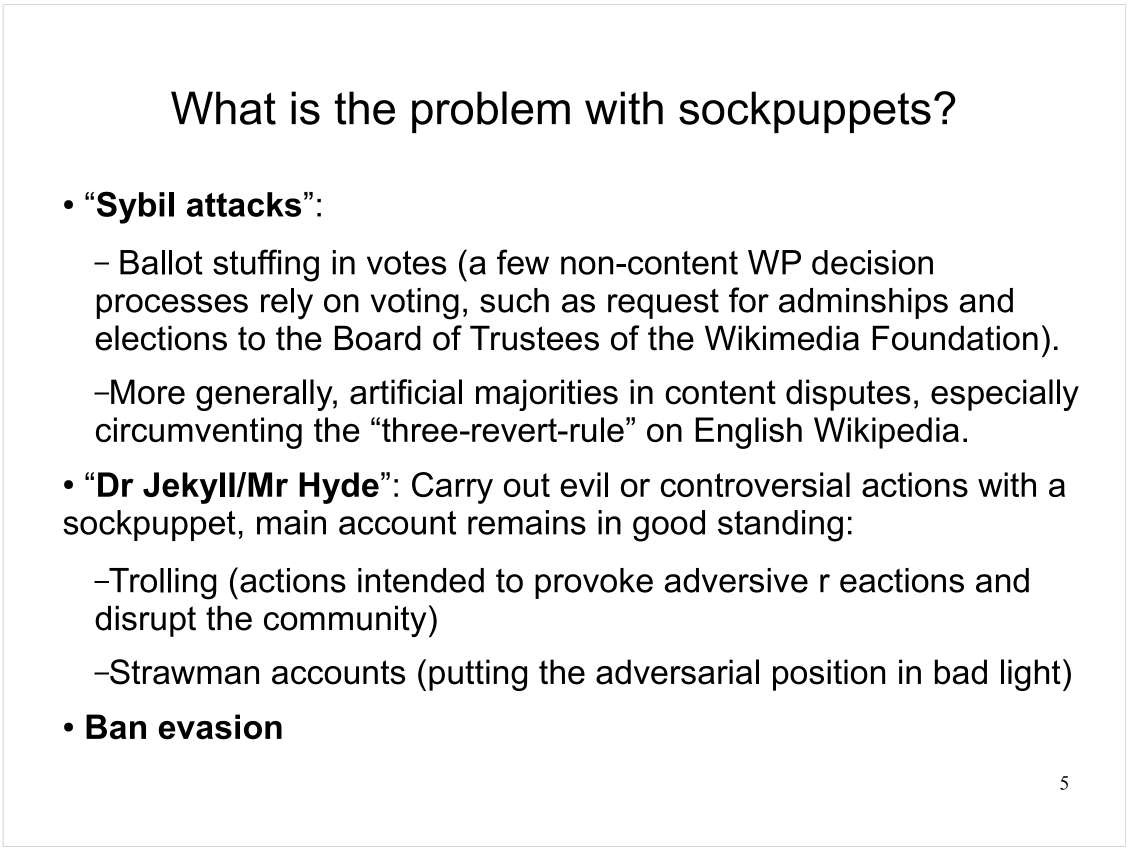
First explain the problem that CU is intended to solve

http://de.wikipedia.org/wiki/Benutzer:Bdk/SPA (“Freiwillige Sockenpuppen-Auskunft”) lists many sockpuppets considered as legitimate on de:WP

See http://commons.wikimedia.org/wiki/Category:Sock_puppets for examples of the Wikipedia community's sockpuppet folklore

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

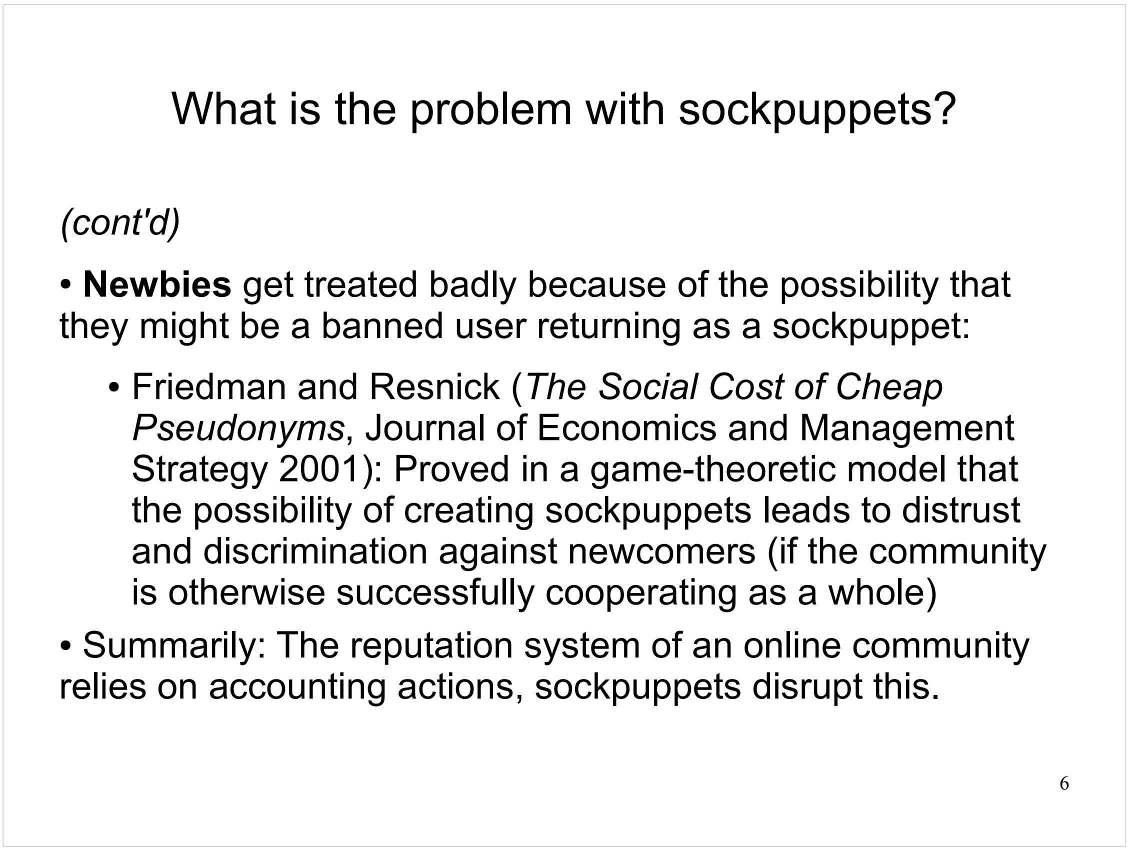
Ventriloquist photo from
http://commons.wikimedia.org/wiki/Image:Mallory_Lewis_and_Lamb_Chop.jpg
(Public Domain)



What is the problem with sockpuppets?

- “**Sybil attacks**”:
  - Ballot stuffing in votes (a few non-content WP decision processes rely on voting, such as request for adminships and elections to the Board of Trustees of the Wikimedia Foundation).
  - More generally, artificial majorities in content disputes, especially circumventing the “three-revert-rule” on English Wikipedia.
- “**Dr Jekyll/Mr Hyde**”: Carry out evil or controversial actions with a sockpuppet, main account remains in good standing:
  - Trolling (actions intended to provoke adversive r eactions and disrupt the community)
  - Strawman accounts (putting the adversarial position in bad light)
- **Ban evasion**

– In principle, content decisions on Wikipedia should be based on consensus, not on votes.
– Real life strawman example: On de:WP, a long time right-wing sockpuppeteer sometimes creates “leftist” sockpuppets.
– “Sybil attack” is an abstract term from theory of social networks and reputation systems



What is the problem with sockpuppets?

(cont'd)

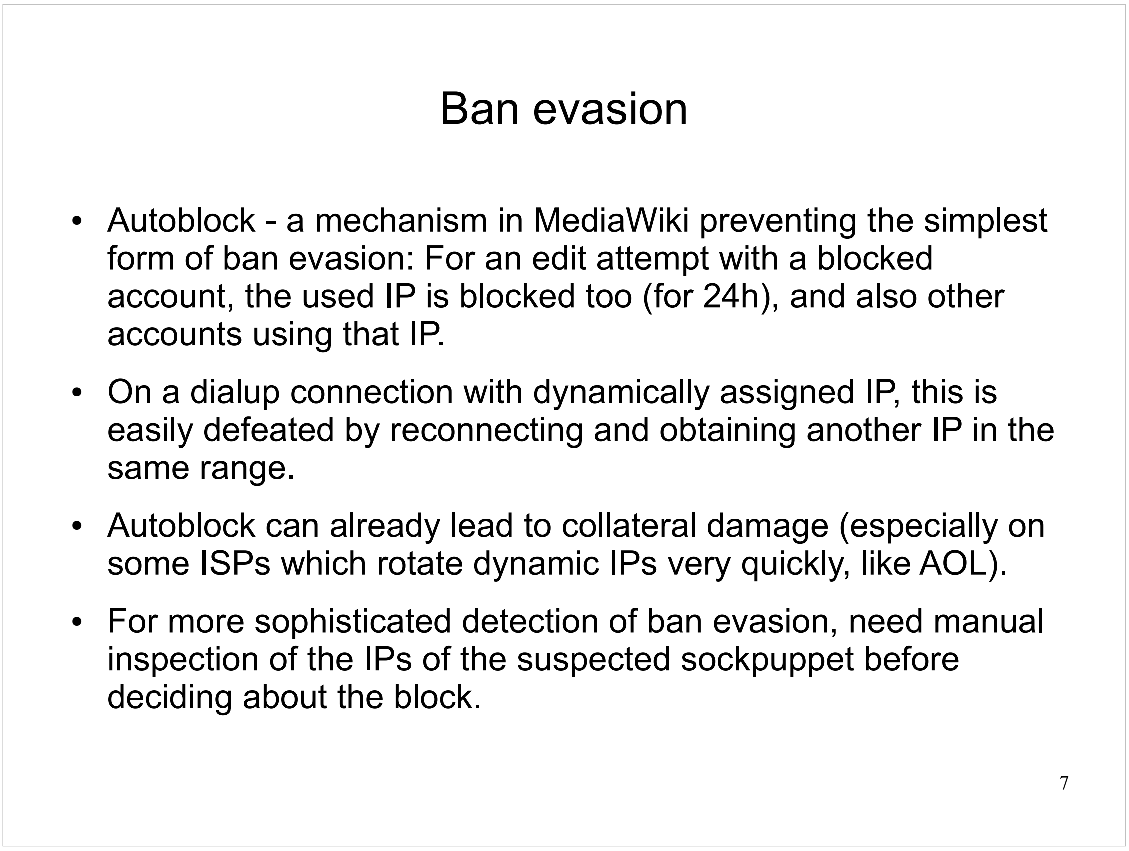
- **Newbies** get treated badly because of the possibility that they might be a banned user returning as a sockpuppet:
    - Friedman and Resnick (*The Social Cost of Cheap Pseudonyms*, Journal of Economics and Management Strategy 2001): Proved in a game-theoretic model that the possibility of creating 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's and Resnick's model assumes that a few malicious players are always present, and that well-meaning players are prone to occasional mistakes.)

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

Shirky (2003) expresses a similar insight in more positive terms: *“...there has to be a penalty for switching handles. .... This keeps the system functioning”* and also says: *”You have to have some cost to either join or participate, if not at the lowest level, then at higher level”*



Ban evasion

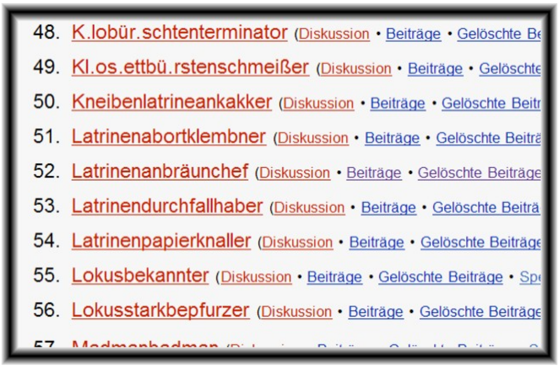
- Autoblock - a mechanism in MediaWiki preventing the simplest form of ban evasion: For an edit attempt with a blocked account, the used IP is blocked too (for 24h), and also other accounts using that IP.
- On a dialup connection with dynamically assigned IP, this is easily defeated by reconnecting and obtaining another IP in the same range.
- Autoblock can already lead to collateral damage (especially on some ISPs which rotate dynamic IPs very quickly, like AOL).
- For more sophisticated detection of ban evasion, need manual inspection of the IPs of the suspected sockpuppet before deciding about the block.

Problem: Can (rarely) lead to “Autoblock cascades”, especially with some ISPs which reassign new IPs quickly → put them on an autoban whitelist

http://en.wikipedia.org/wiki/Wikipedia:Autoblock

Range blocks

- For severe cases of mass vandalism, blocking the whole range is an option.
- Problem: Potentially huge collateral damage. Can be estimated by looking at previous “good” edits from that range. If a hard block is considered, logged-in edits are of interest too - needs a range CU check, problematic



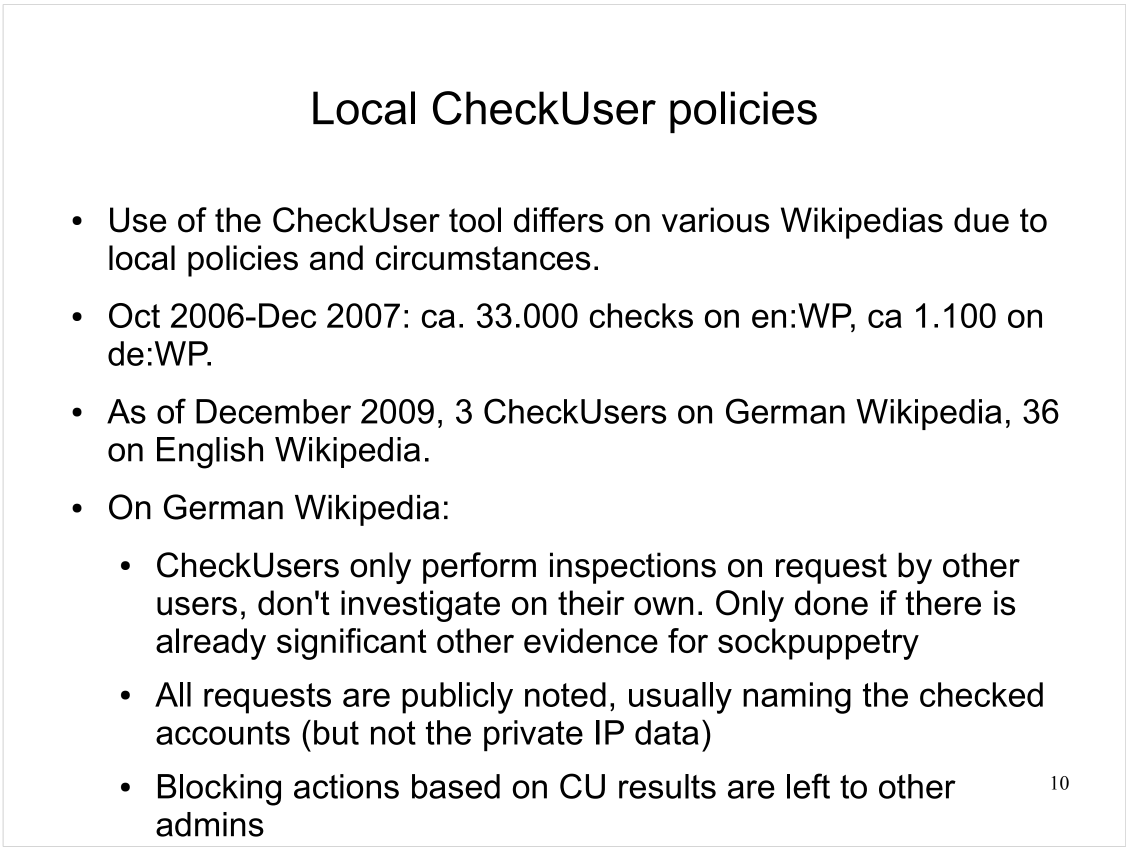
48. K.lobür.schtenterminator (Diskussion • Beiträge • Gelöschte Be
49. Kl.os.ettbü.rstenschmeißer (Diskussion • Beiträge • Gelöschte
50. Kneibenlatrineankakker (Diskussion • Beiträge • Gelöschte Beitr
51. Latrinenabortklembner (Diskussion • Beiträge • Gelöschte Beiträ
52. Latrinenanbräunchef (Diskussion • Beiträge • Gelöschte Beiträge
53. Latrinendurchfallhaber (Diskussion • Beiträge • Gelöschte Beiträ
54. Latrinenpapierknaller (Diskussion • Beiträge • Gelöschte Beiträge
55. Lokusbekannter (Diskussion • Beiträge • Gelöschte Beiträge • Spe
56. Lokusstarkbepfurzer (Diskussion • Beiträge • Gelöschte Beiträge

Excerpt from a list of many thousand sock puppets created on de:WP by a vandal with a toilet fixation (since at least 2006, still active – one of the few cases where range blocks are used regularly)

Apologies to the German speakers, I hope everyone had their lunch already

The Checkuser tool in MediaWiki

- Allows a few trusted users on a wiki to manually inspect IP addresses from which edits are made from (not: reader's IPs)
- On many web sites and web forums, administrators can routinely see participant's IPs, email adresses etc.
- Wikimedia Foundation's privacy policy sets higher expectations.
- Access to the CheckUser tool is only granted to a few trusted users on each wiki. Must be over 18 years and identify themselves to the Wikimedia Foundation
- Each access is logged (but log is only visible to other CheckUsers due to privacy concerns)
- Ombudsman commission for investigation of alleged violations of the privacy policy



Local CheckUser policies

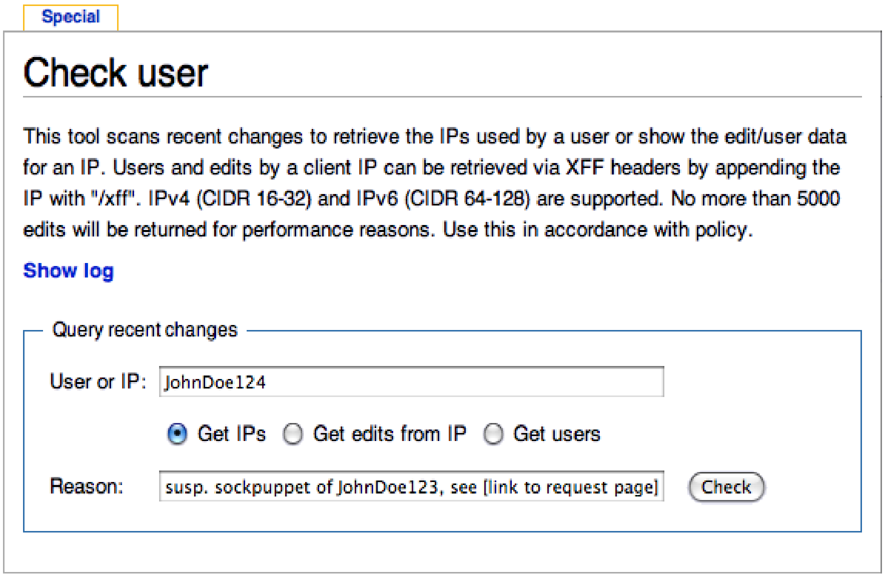
- Use of the CheckUser tool differs on various Wikipedias due to local policies and circumstances.
- Oct 2006-Dec 2007: ca. 33.000 checks on en:WP, ca 1.100 on de:WP.
- As of December 2009, 3 CheckUsers on German Wikipedia, 36 on English Wikipedia.
- On German Wikipedia:
  - CheckUsers only perform inspections on request by other users, don't investigate on their own. Only done if there is already significant other evidence for sockpuppetry
  - All requests are publicly noted, usually naming the checked accounts (but not the private IP data)
  - Blocking actions based on CU results are left to other admins

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

https://secure.wikimedia.org/wikipedia/de/wiki/Wikipedia:CUA

I'm one of these three.

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]

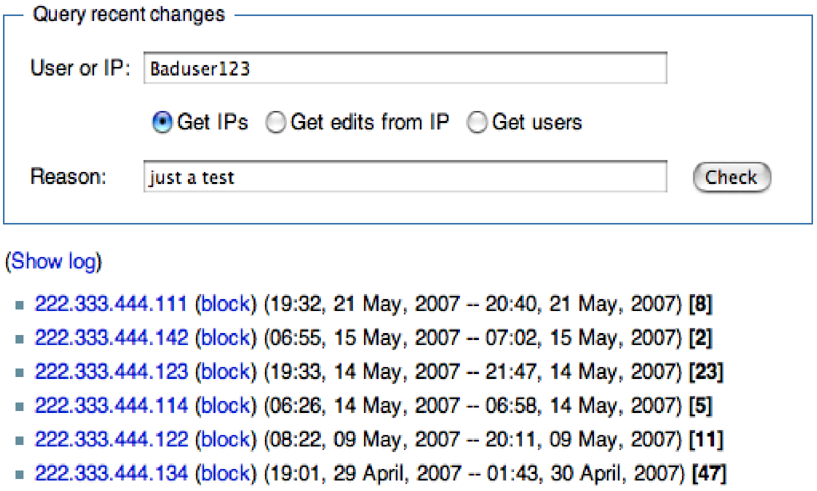
"Get IPs": Retrieve IP addresses from which this user account has edited

"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



Query recent changes

User or IP: Baduser123

● Get IPs ○ Get edits from IP ○ Get users

Reason: just a test

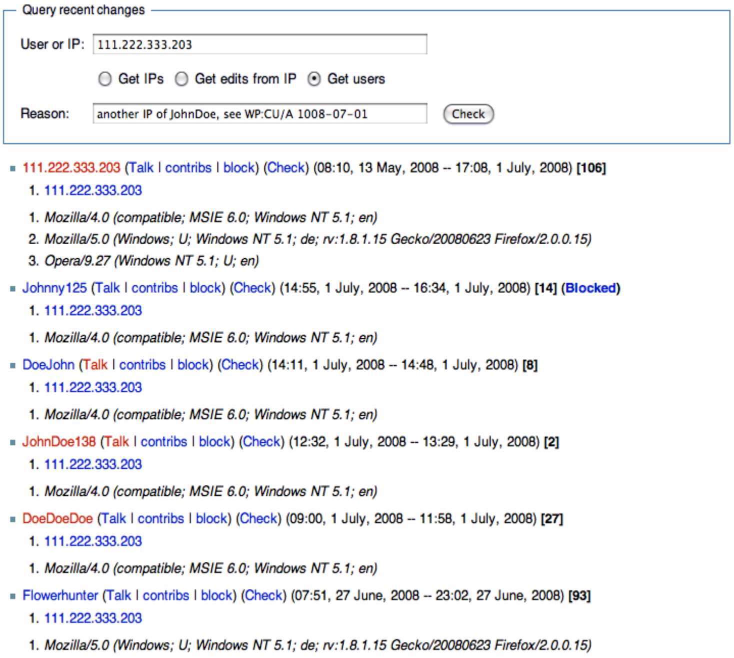
(Show log)

- 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]**

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 users editing from an IP



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

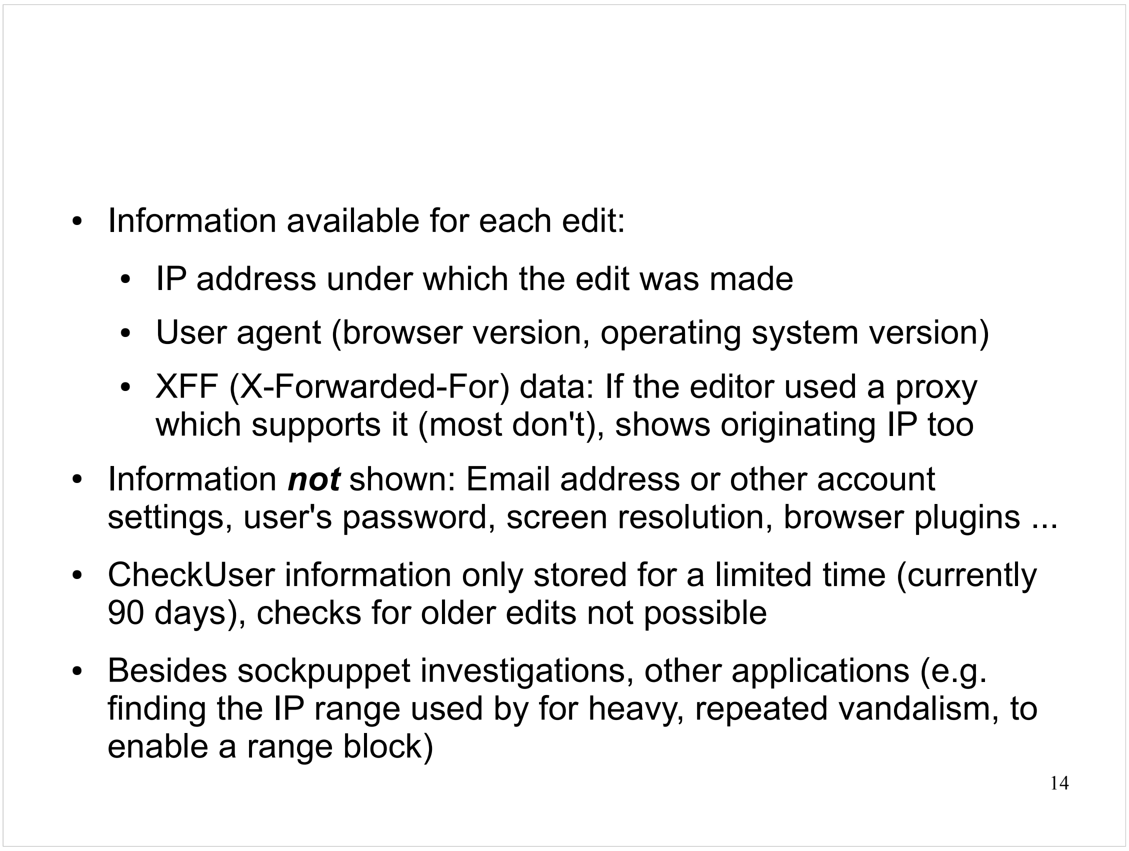
  1. *Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; en)*
- DoeDoeDoe (Talk | contribs | block) (Check) (09:00, 1 July, 2008 -- 11:58, 1 July, 2008) **[27]**
  1. 111.222.333.203

  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)*

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)

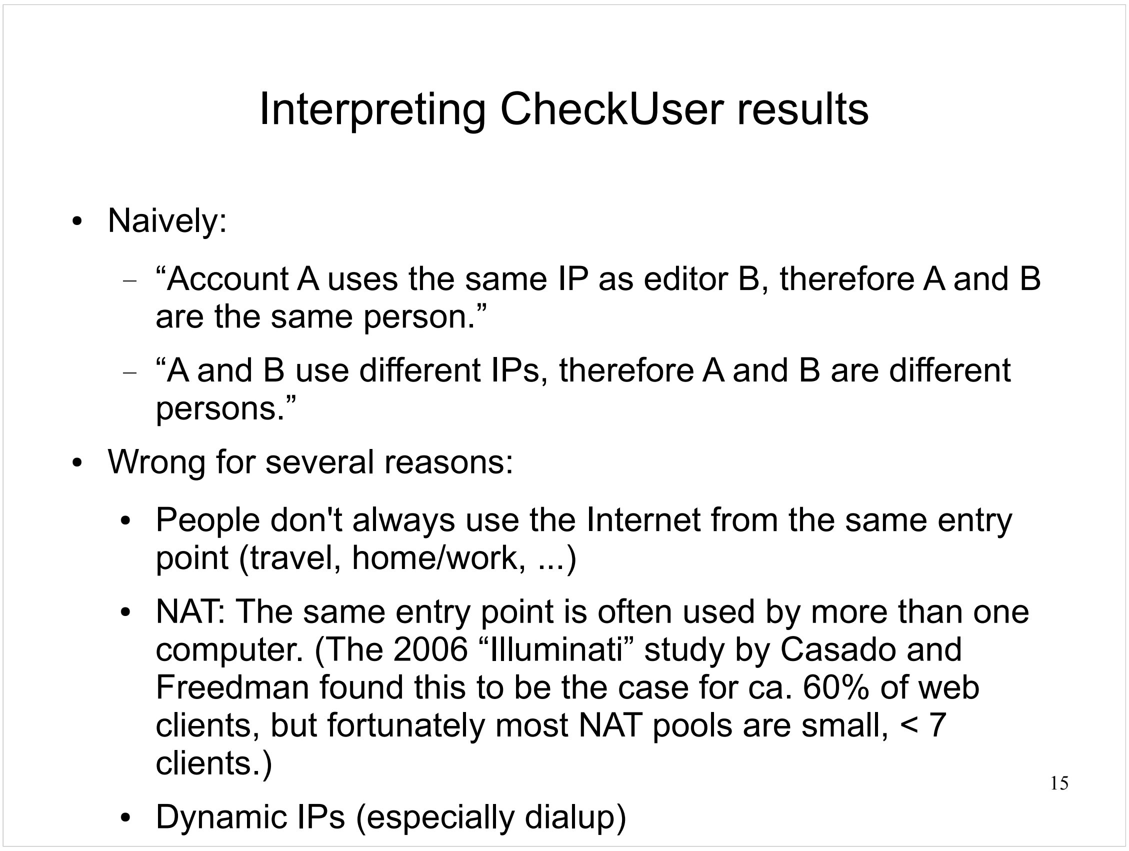
(Some browser add-ons do show up in the user agent, though.)

Statistics helping to estimate the frequency of a particular user agent on Wikipedia (from readers, not editors, though):

http://stats.wikimedia.org/wikimedia/squids/SquidReportOperatingSystems.htm
http://stats.wikimedia.org/wikimedia/squids/SquidReportClients.htm

Frequent suggestion addressing the privacy concerns: Instead of the actual IP, just store some kind of hash. But this would loose much important information which is used regularly in sockpuppet investigations, for example WHOIS data, geolocation, whether two IPs were in the same dynamic range, etc.

See slide 8 about range blocks



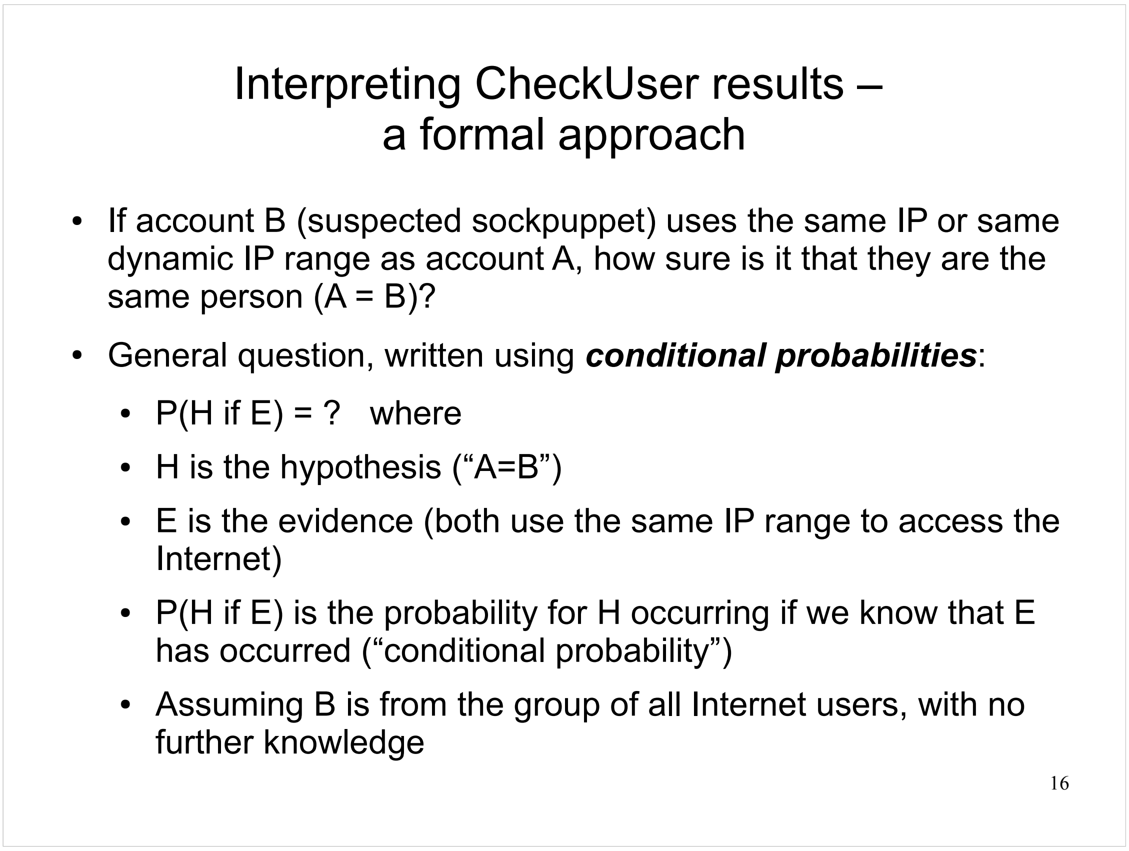
Interpreting CheckUser results

- 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, ...)
  - NAT: The same entry point is often used by more than one computer. (The 2006 “Illuminati” study by Casado and Freedman found this to be the case for ca. 60% of web clients, but fortunately most NAT pools are small, < 7 clients.)
  - Dynamic IPs (especially dialup)

On this and the following slides, user agents and XFF are ignored for simplicity.

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”).

http://illuminati.coralcdn.org/
They also studied how fast dynamic IPs are usually rotated (generally “on the order of several days”)



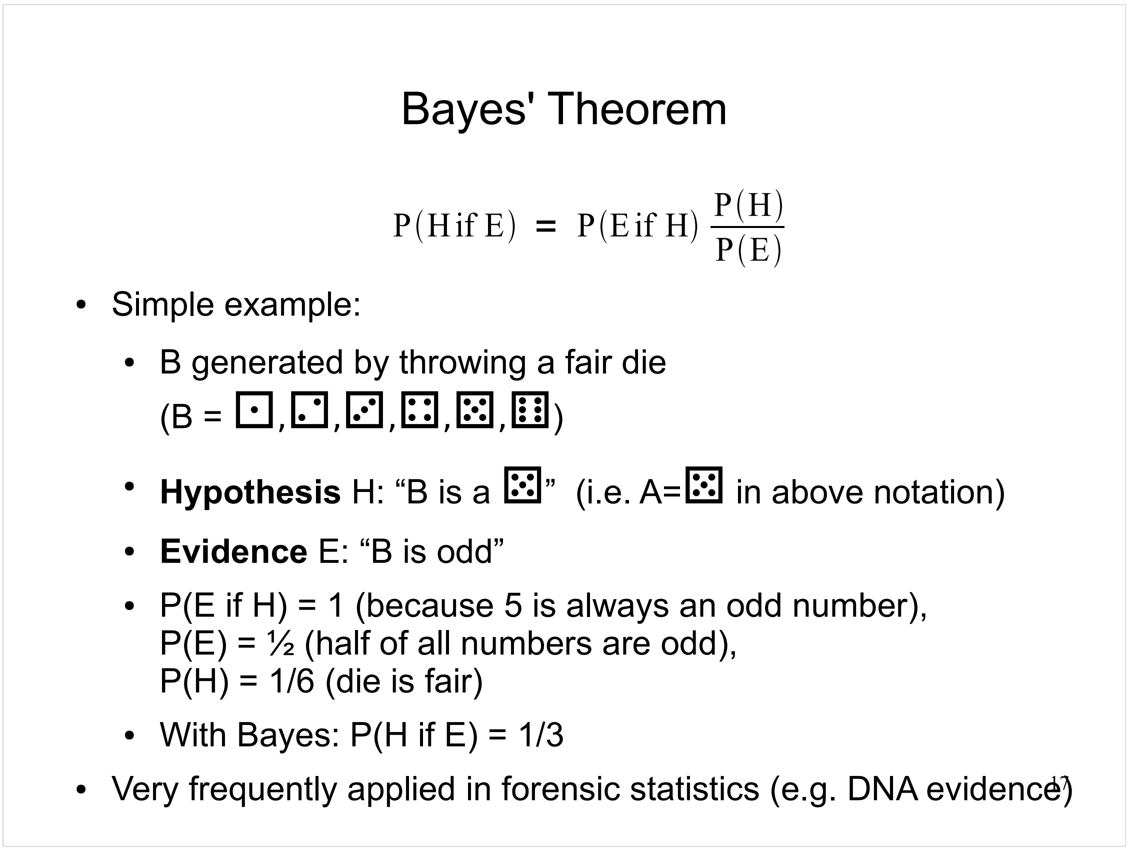
Interpreting CheckUser results – a formal approach

- 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 if E) = ?   where
  - H is the hypothesis (“A=B”)
  - E is the evidence (both use the same IP range to access the Internet)
  - P(H if 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.

Usual notation is “P(H|E)” instead of “P(H if E)”



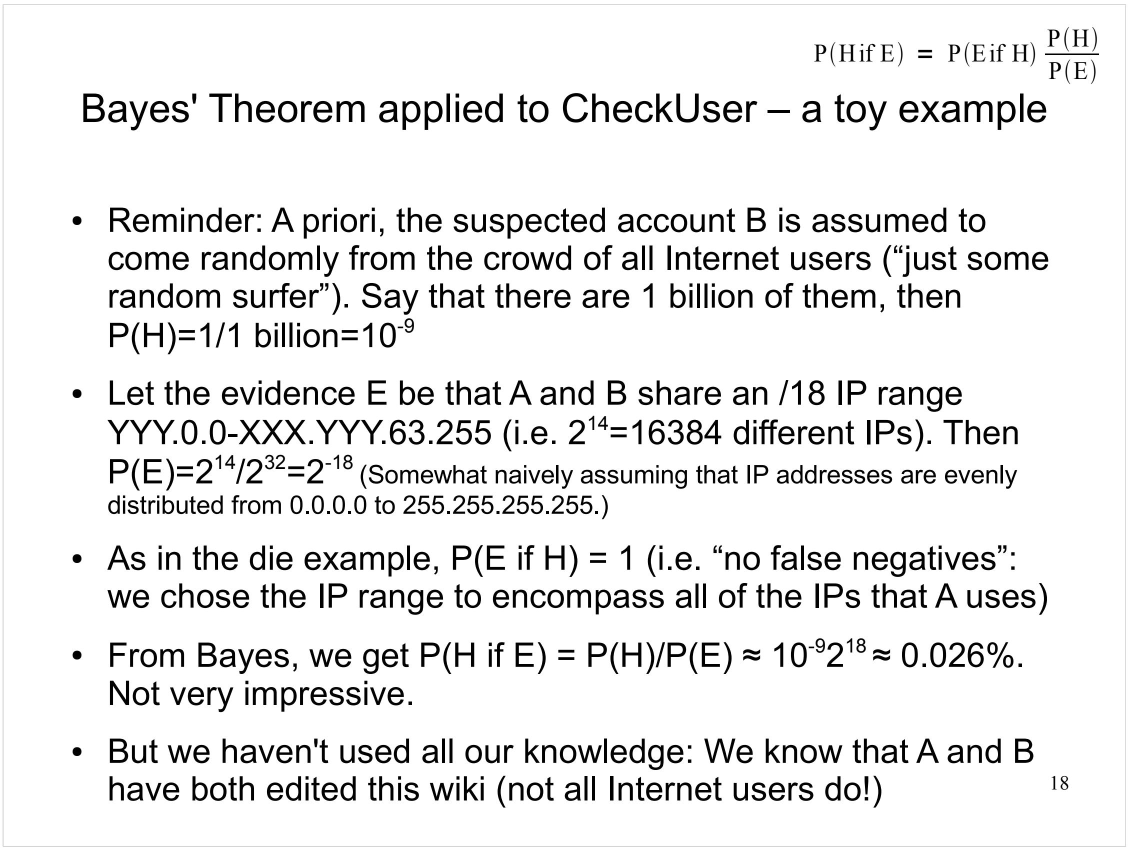
Bayes' Theorem

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

- 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 if 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 if E) = 1/3
- Very frequently applied in forensic statistics (e.g. DNA evidence)

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

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



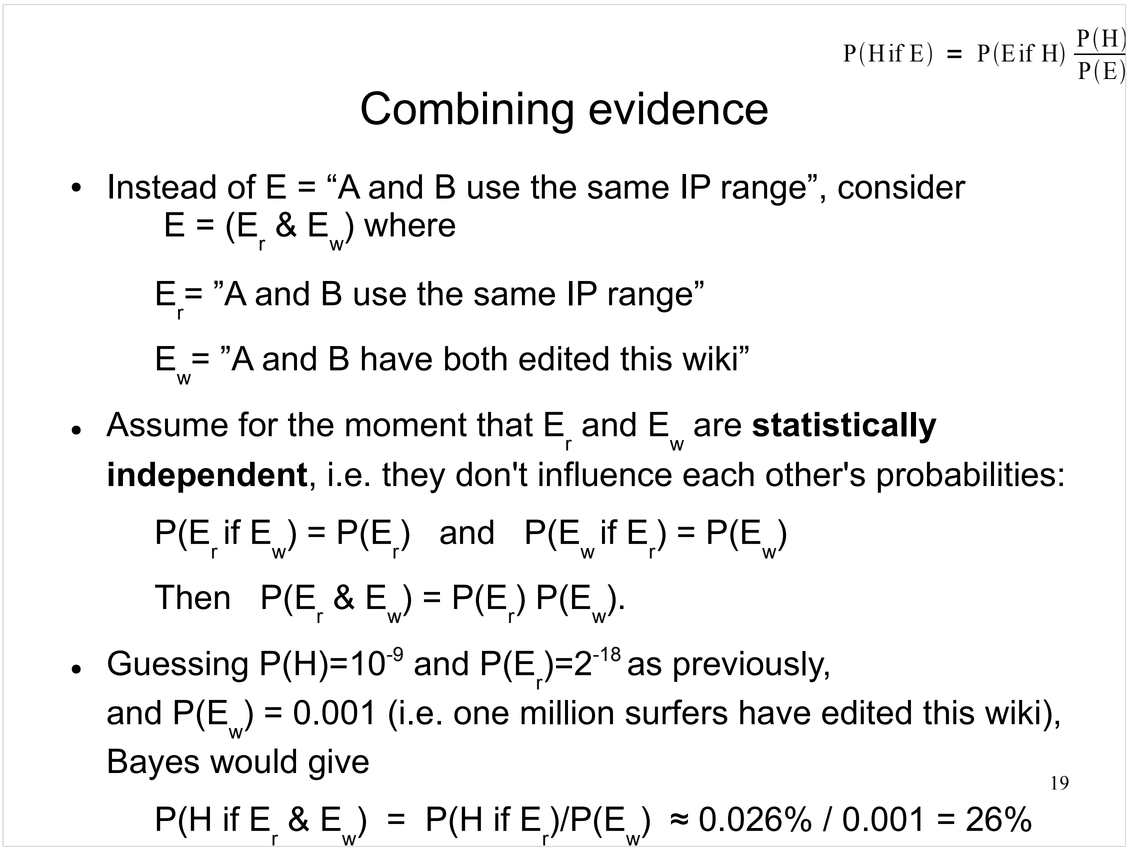
$$P(H \text{ if } E) = P(E \text{ if } H)\,\frac{P(H)}{P(E)}$$

Bayes' Theorem applied to CheckUser – a toy example

- Reminder: A priori, the suspected account 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}
- Let the evidence E be that A and B share an /18 IP range YYY.0.0-XXX.YYY.63.255 (i.e. 2^{14}=16384 different IPs). Then P(E)=$2^{14}/2^{32}=2^{-18}$ (Somewhat naively assuming that IP addresses are evenly distributed from 0.0.0.0 to 255.255.255.255.)
- As in the die example, P(E if 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 if 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”



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

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:

  $P(E_r \text{ if } E_w) = P(E_r)$ and $P(E_w \text{ if } E_r) = P(E_w)$

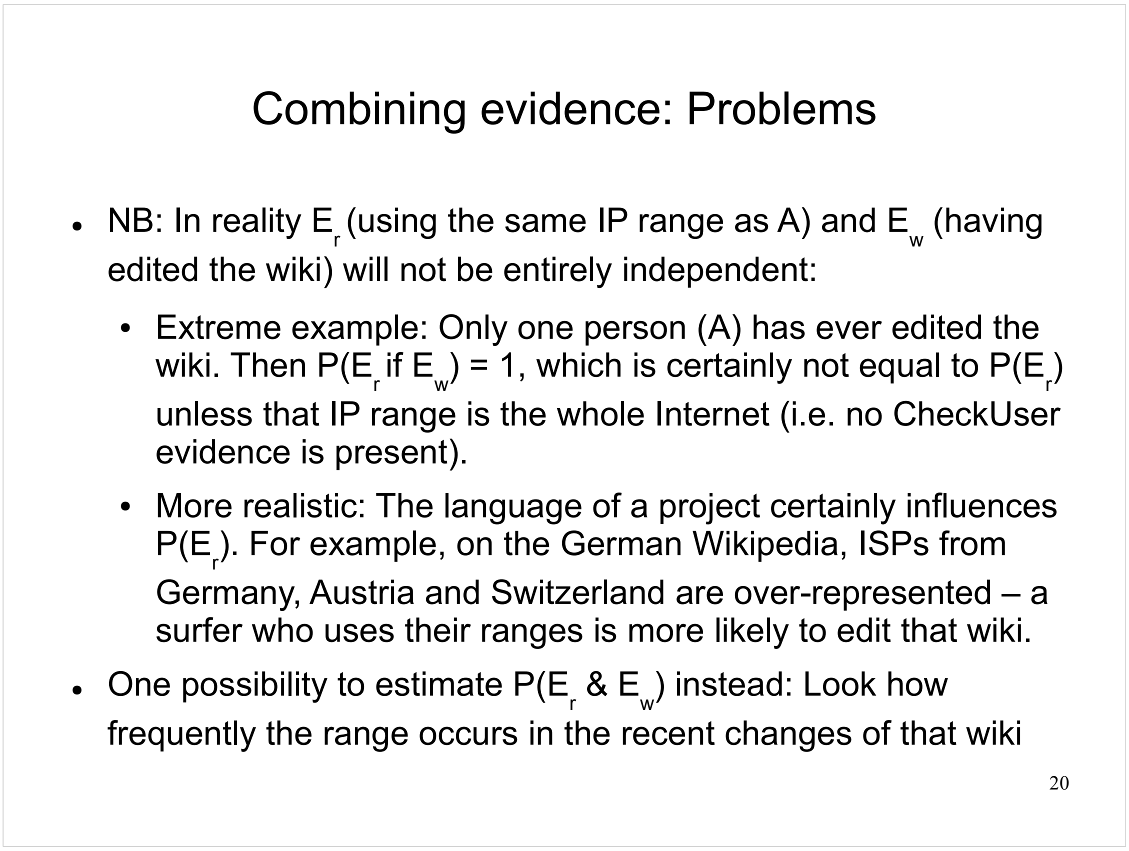
  Then $P(E_r \,\&\, E_w) = P(E_r)\, P(E_w)$.
- Guessing $P(H)=10^{-9}$ and $P(E_r)=2^{-18}$ as previously,
  and $P(E_w) = 0.001$ (i.e. one million surfers have edited this wiki), Bayes would give

  $P(H \text{ if } E_r \,\&\, E_w) = P(H \text{ if } E_r)/P(E_w) \approx 0.026\% / 0.001 = 26\%$

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.



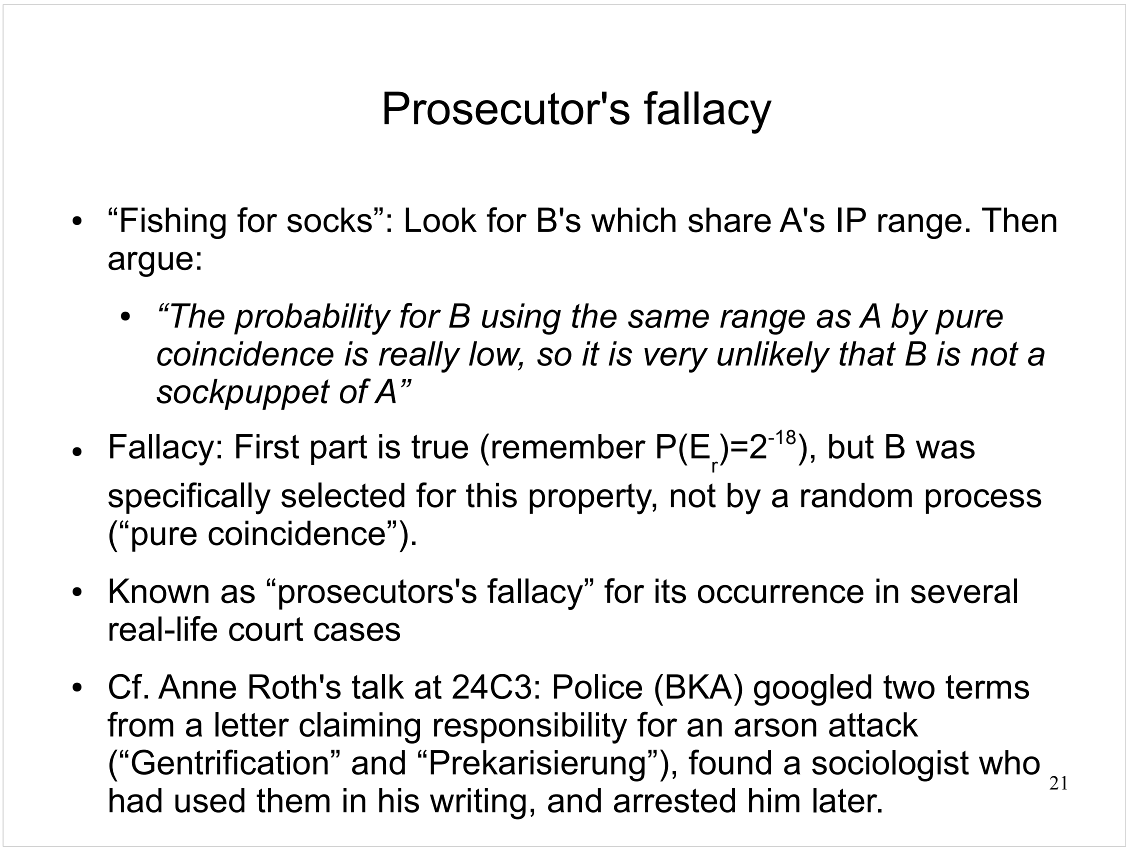
Combining evidence: Problems

- NB: In reality E_r (using the same IP range as 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 $P(E_r \text{ if } E_w) = 1$, 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 – a surfer who uses their ranges is more likely to edit that wiki.
- 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 “a priori” 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
- Cf. Anne Roth's talk at 24C3: Police (BKA) googled two terms from a letter claiming responsibility for an arson attack (“Gentrification” and “Prekarisierung”), found a sociologist who had used them in his writing, and arrested him later.

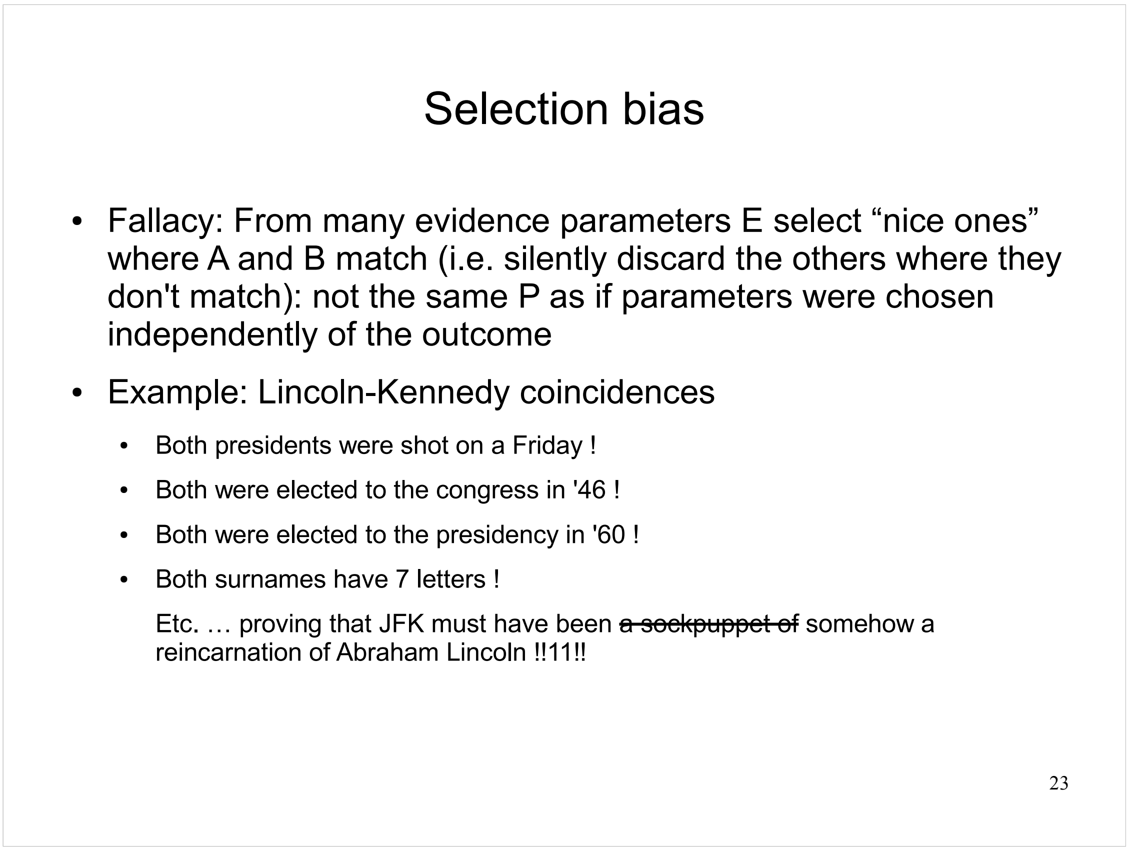
Argument basically claims “P(E if (not H)) = 1 - P(H if E)”, but this equality is not valid

http://www.taz.de/index.php?id=start&art=3471&id=deutschland-artikel&cHash=5218eee73a

To be fair to the BKA, there seems to have been some other evidence, but it does not appear to have been statistically independent (people who meet over similar political views are likely to use terms associated with these views).

Combining with non-CU evidence, defendant's fallacy

- Recall that in our numerical example, P(“A = B” if E) was still small (nowhere near 1), even when combining IP range and being an editor as evidence E
- Other evidence from CheckUser results: 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. Wikipedia's radical transparency means that a lot of such evidence is available from user contribution list, see next slides.
- “**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.



Selection bias

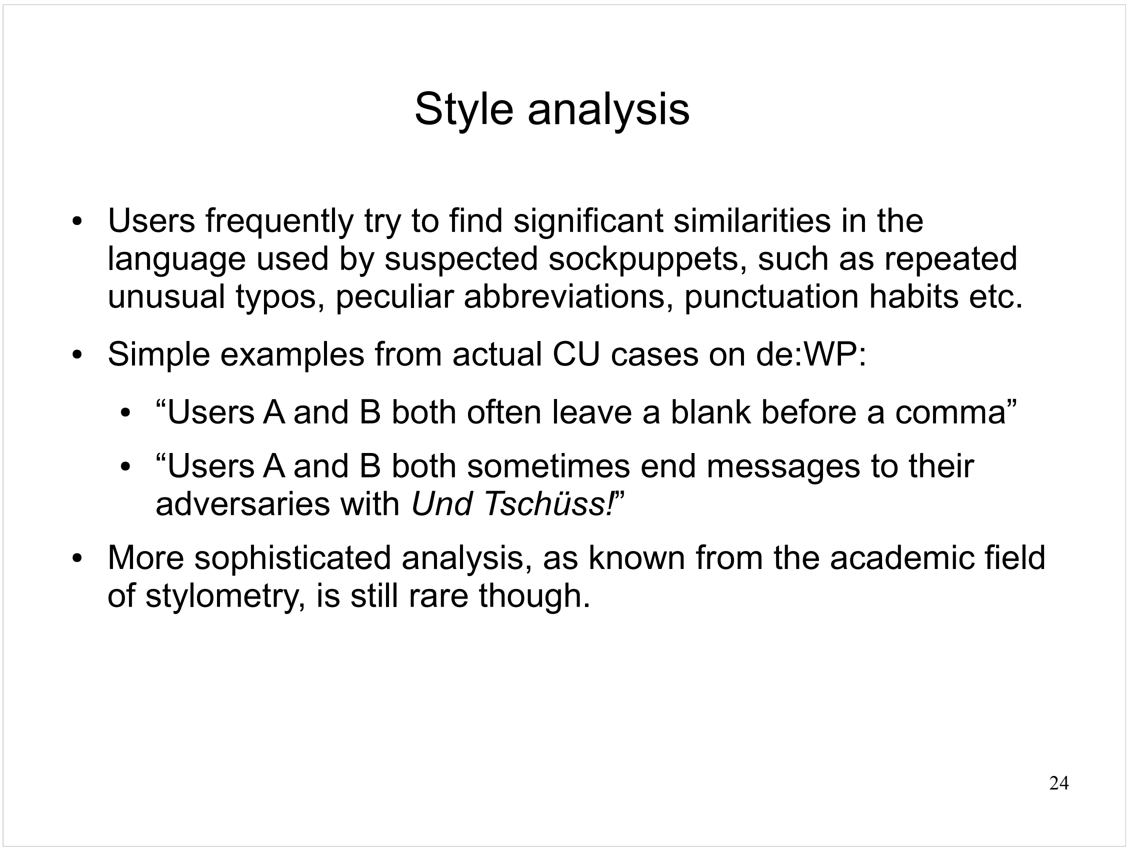
- Fallacy: From many evidence parameters E select “nice ones” where A and B match (i.e. silently discard the others where they don't match): not the same P as if parameters were chosen independently of the outcome
- Example: Lincoln-Kennedy coincidences
  - Both presidents were shot on a Friday !
  -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

  Etc. … proving that JFK must have been ~~a sockpuppet of~~ somehow a reincarnation of Abraham Lincoln !!11!!

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



Style analysis

- Users frequently try to find significant similarities in the language used by suspected sockpuppets, such as repeated unusual typos, peculiar abbreviations, punctuation habits etc.
- Simple examples from actual CU cases on de:WP:
  - “Users A and B both often leave a blank before a comma”
  - “Users A and B both sometimes end messages to their adversaries with *Und Tschüss!*”
- More sophisticated analysis, as known from the academic field of stylometry, is still rare though.

https://secure.wikimedia.org/wikipedia/de/wiki/Wikipedia:Checkuser/Anfragen/Archiv/2008-2#.2813._Mai.29_-_Trintheim_und_Computare

Stylometry, forensic linguistics

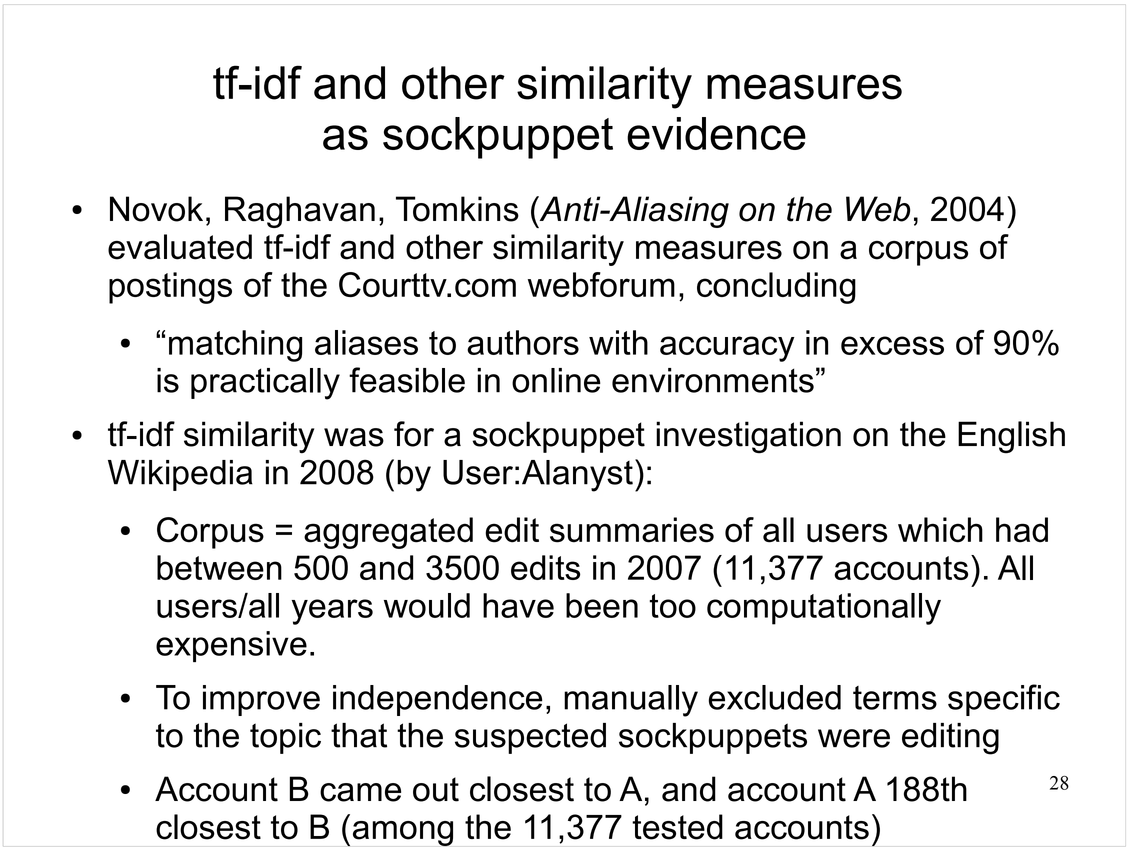
- History: Attempts to determine authorship of Shakespeare's works, the Federalist papers, the Unabomber manifesto...
- Underlying assumption: While people vary their writing style according to occasion, genre, mood etc., there exist persistent habits and traits which distinguish individual writers.
- Which properties can be regarded as persistent is often controversial, but the field has 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. (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 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 too 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. Next slide) 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 full edit histories of all users on en:WP, 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:

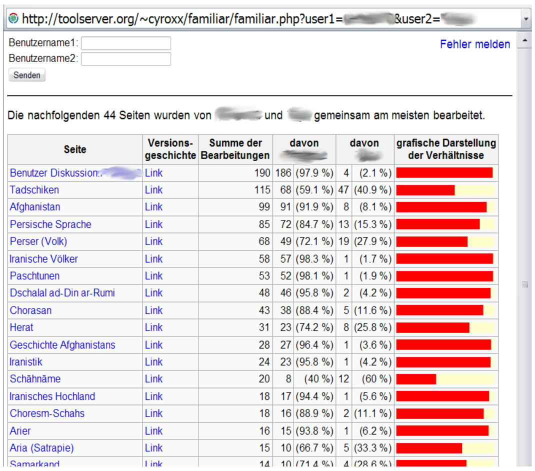
http://citeseerx.ist.psu.edu/viewdoc/download;?doi=10.1.1.2.3205&rep=rep1&type=pdf

Sockpuppet investigation by Alanyst:

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

Similar interests

- Just a few personal interests and cultural preferences can suffice to identify an individual (cf. 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)”



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.

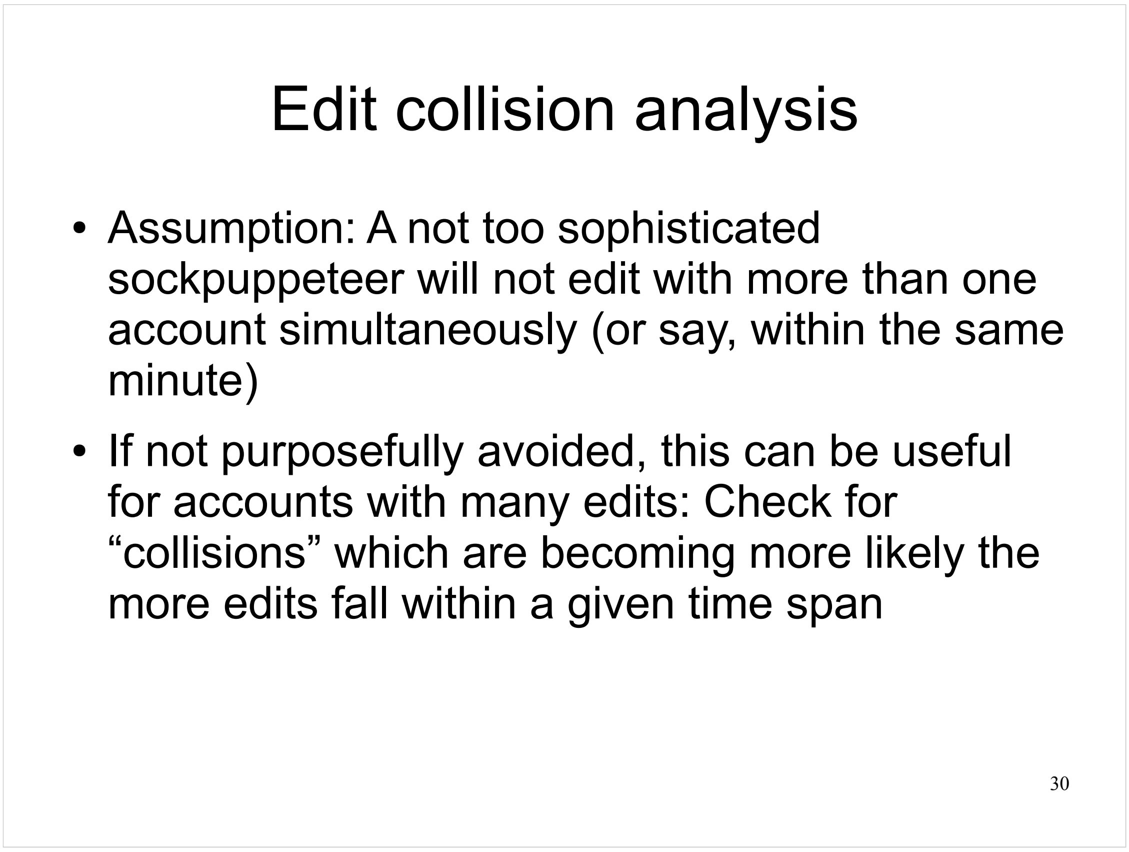
Seite	Versions-geschichte	Summe der Bearbeitungen	davon [redacted]		davon [redacted]		grafische Darstellung der Verhältnisse
Benutzer Diskussion:[redacted]	Link	190	186	(97.9 %)	4	(2.1 %)	🟥
Tadschiken	Link	115	68	(59.1 %)	47	(40.9 %)	🟥
Afghanistan	Link	99	91	(91.9 %)	8	(8.1 %)	🟥
Persische Sprache	Link	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 %)	🟥

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

Due to the universal scope of an encyclopedia, one is more likely to reveal several unrelated areas of personal interest and knowledge on Wikipedia than, say, on a typical web forum.

Screenshot from http://toolserver.org/~cyroxx/familiar/familiar.php (tool by user CyRoXX, currently only available on the German Wikipedia. Similar tool for English Wikipedia: http://toolserver.org/~mzmcbride/cgi-bin/wikistalk.py)

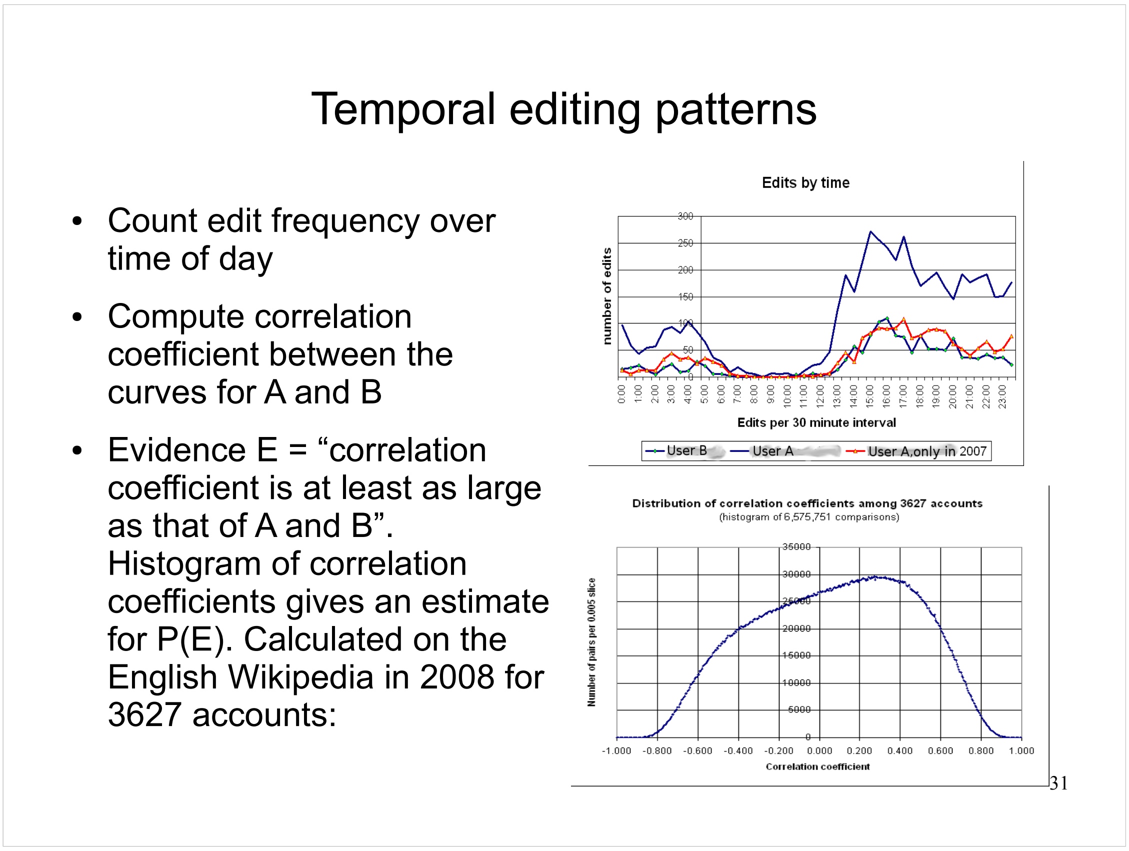


Edit collision analysis

- Assumption: A not too sophisticated sockpuppeteer will not edit with more than one account simultaneously (or say, within the same minute)
- If not purposefully avoided, this can be useful for accounts with many edits: Check for “collisions” which are becoming more likely the more edits fall within a given time span

Tool which is used to detect collisions easily:
http://toolserver.org/~erwin85/contribs.php

See also http://en.wikipedia.org/wiki/User:Alanyst/Edit_collision_research

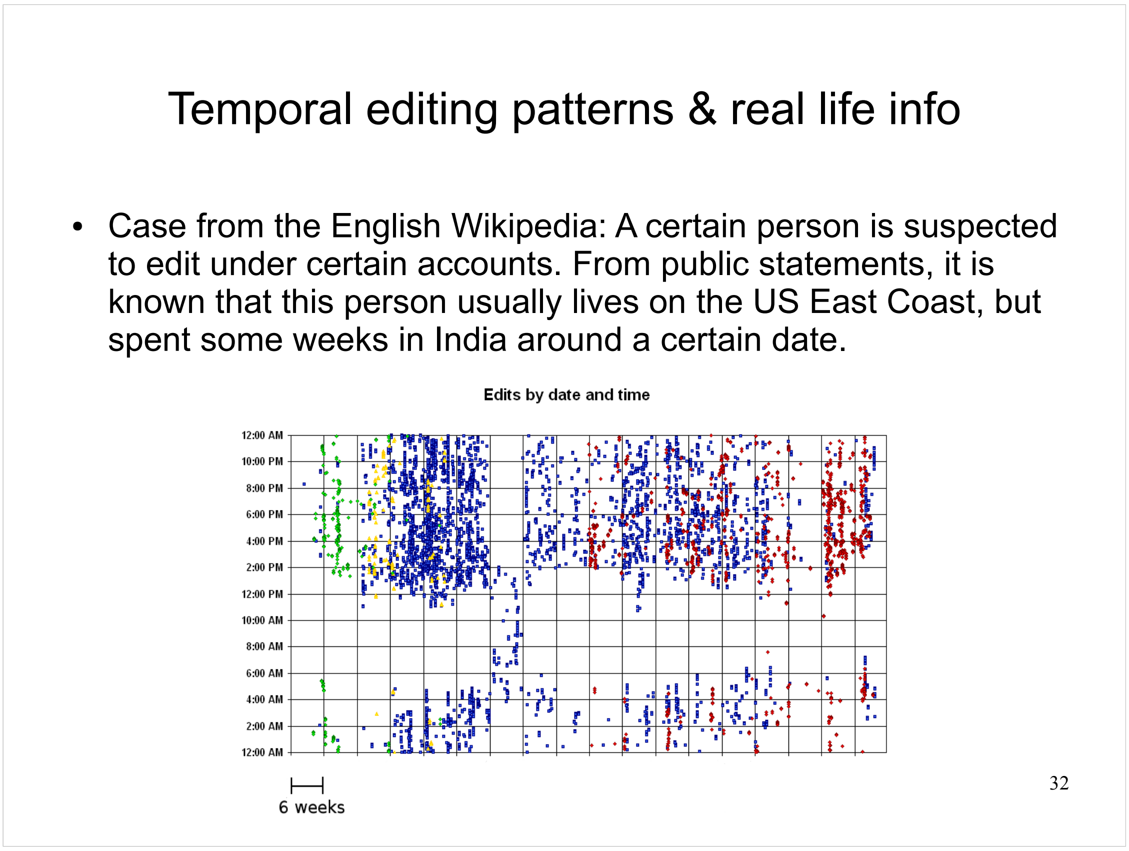


Temporal editing patterns

- Count edit frequency over time of day
- Compute correlation coefficient between the curves for A and B
- Evidence E = “correlation coefficient is at least as large as that of A and B”. Histogram of correlation coefficients gives an estimate for P(E). Calculated on the English Wikipedia in 2008 for 3627 accounts:

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

Other Wikipedians have calculated such correlation coefficients too (also came up as evidence on de:WP), but afaik no publicly available tool



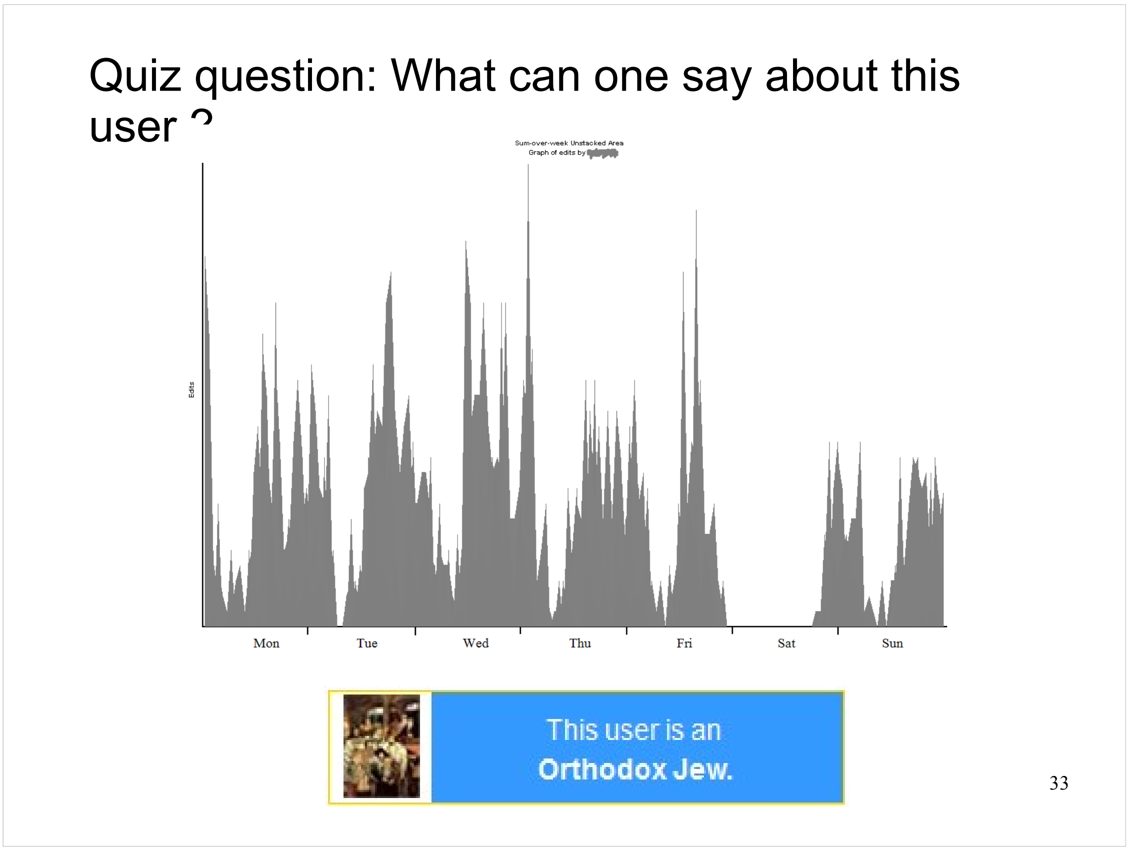
Temporal editing patterns & real life info

- 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.

Edits by date and time

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



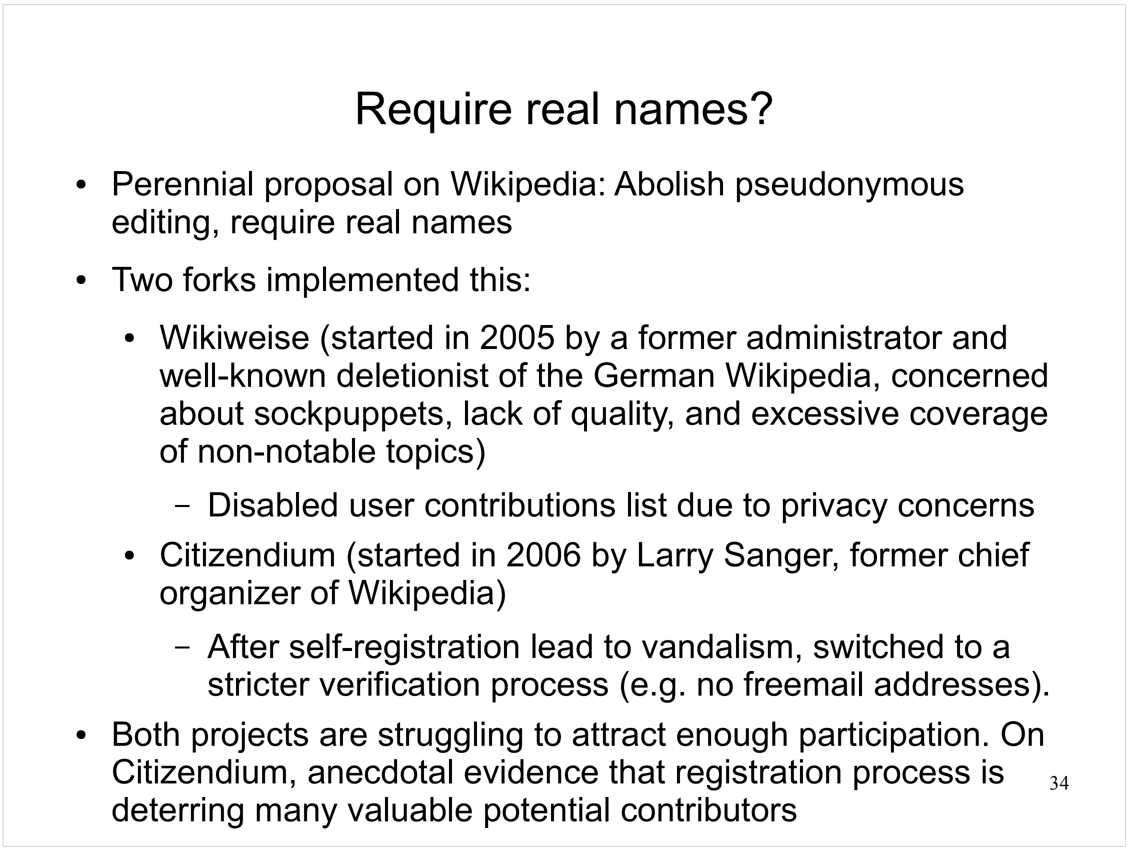
Quiz question: What can one say about this user ?

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 FIcelloguy's Tool: http://en.wikipedia.org/wiki/Wikipedia:WPEC/FT/H
(currently broken)

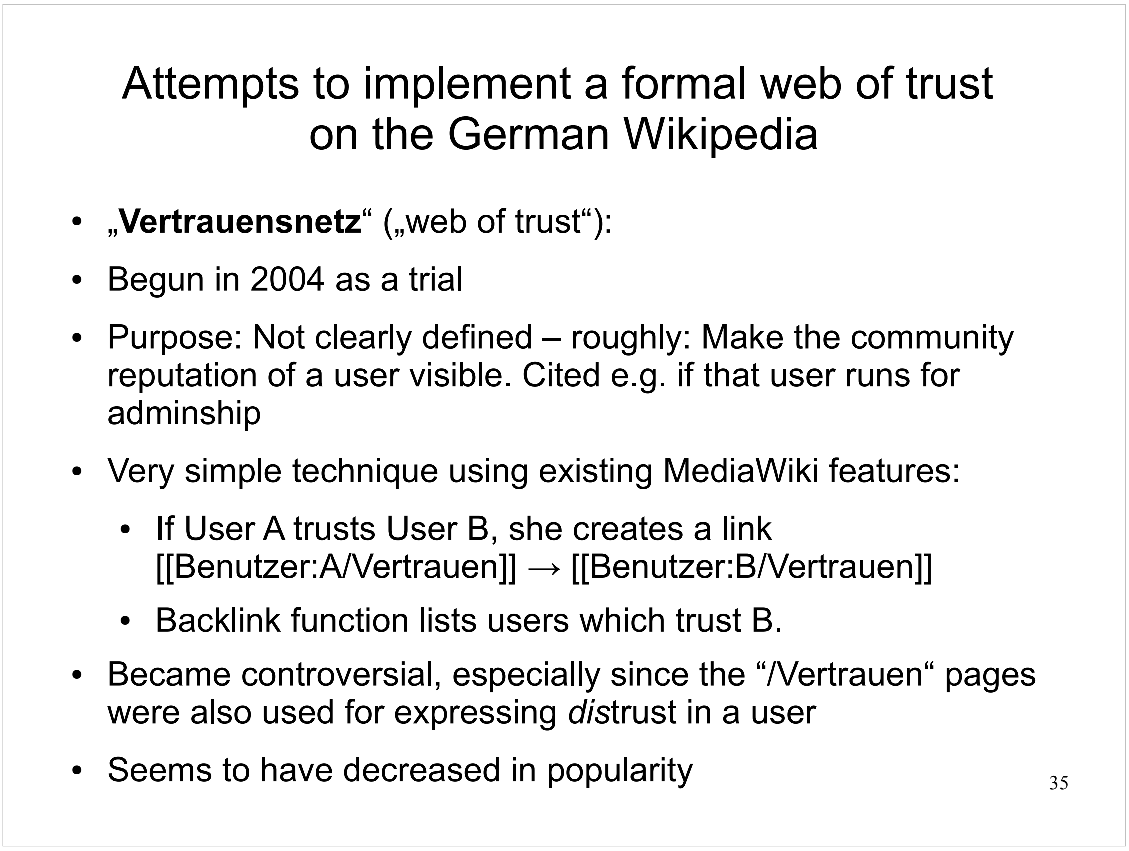


Require real names?

- Perennial proposal on Wikipedia: Abolish pseudonymous editing, require real names
- Two forks implemented this:
  - Wikiweise (started in 2005 by a former administrator and well-known deletionist of the German Wikipedia, concerned about sockpuppets, lack of quality, and excessive coverage of non-notable topics)
    - Disabled user contributions list due to privacy concerns
  - Citizendium (started in 2006 by Larry Sanger, former chief organizer of Wikipedia)
    - After self-registration lead to vandalism, switched to a stricter verification process (e.g. no freemail addresses).
- Both projects are struggling to attract enough participation. On Citizendium, anecdotal evidence that registration process is deterring many valuable potential contributors

See http://www.wikiweise.de/wiki/Wikiweise%3AWikiweise%20und%20Wikipedia

See the notes for my talk about Citizendium at Wikimania 2009: http://commons.wikimedia.org/wiki/File:Lessons_from_Citizendium.pdf (also for some numbers on Wikiweise)

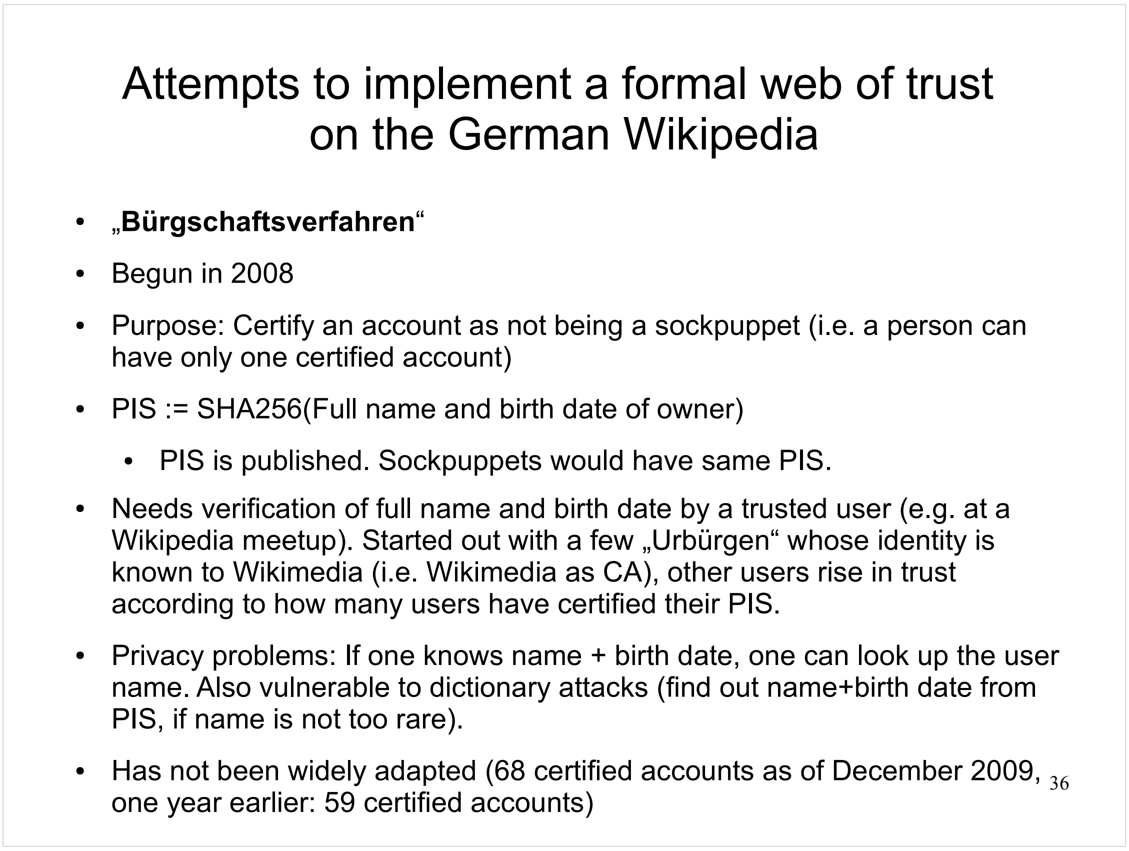


Attempts to implement a formal web of trust on the German Wikipedia

- „**Vertrauensnetz**“ („web of trust“):
- Begun in 2004 as a trial
- Purpose: Not clearly defined – roughly: Make the community reputation of a user visible. Cited e.g. if that user runs for adminship
- Very simple technique using existing MediaWiki features:
  - If User A trusts User B, she creates a link [[Benutzer:A/Vertrauen]] → [[Benutzer:B/Vertrauen]]
  - Backlink function lists users which trust B.
- Became controversial, especially since the “/Vertrauen“ pages were also used for expressing *dis*trust in a user
- Seems to have decreased in popularity

http://de.wikipedia.org/wiki/Wikipedia:Vertrauensnetz

http://de.wikipedia.org/wiki/Wikipedia:Umfragen/Vertrauensnetz (2008)



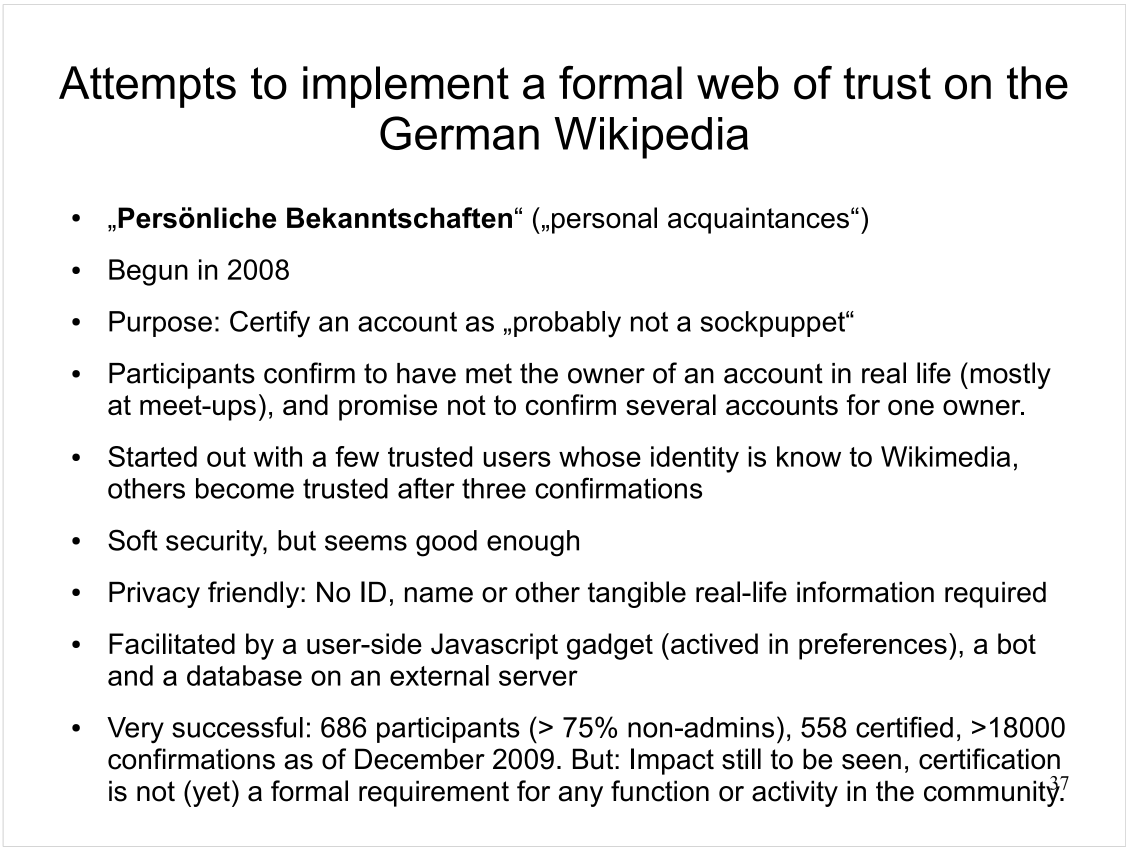
Attempts to implement a formal web of trust on the German Wikipedia

- **„Bürgschaftsverfahren“**
- Begun in 2008
- Purpose: Certify an account as not being a sockpuppet (i.e. a person can have only one certified account)
- PIS := SHA256(Full name and birth date of owner)
  - PIS is published. Sockpuppets would have same PIS.
- Needs verification of full name and birth date by a trusted user (e.g. at a Wikipedia meetup). Started out with a few „Urbürgen“ whose identity is known to Wikimedia (i.e. Wikimedia as CA), other users rise in trust according to how many users have certified their PIS.
- Privacy problems: If one knows name + birth date, one can look up the user name. Also vulnerable to dictionary attacks (find out name+birth date from PIS, if name is not too rare).
- Has not been widely adapted (68 certified accounts as of December 2009, one year earlier: 59 certified accounts)

http://de.wikipedia.org/wiki/Benutzer:YourEyesOnly/Bürgschaft

Strong security approach, not unlike CACert or other cryptographic webs of trust, meetups as keysigning parties

Of course, an owner of a certified account can still have other, un-certified accounts.



Attempts to implement a formal web of trust on the German Wikipedia

- „**Persönliche Bekanntschaften**“ („personal acquaintances“)
- Begun in 2008
- Purpose: Certify an account as „probably not a sockpuppet“
- Participants confirm to have met the owner of an account in real life (mostly at meet-ups), and promise not to confirm several accounts for one owner.
- Started out with a few trusted users whose identity is know to Wikimedia, others become trusted after three confirmations
- Soft security, but seems good enough
- Privacy friendly: No ID, name or other tangible real-life information required
- Facilitated by a user-side Javascript gadget (actived in preferences), a bot and a database on an external server
- Very successful: 686 participants (> 75% non-admins), 558 certified, >18000 confirmations as of December 2009. But: Impact still to be seen, certification is not (yet) a formal requirement for any function or activity in the community.

http://de.wikipedia.org/wiki/Wikipedia:Persönliche_Bekanntschaften

Concerns:
- There are users who are unable or unwilling to attend meetups
- Might work less well on geographically more dispersed projects like en:WP
- Might foster old boys' networks

Similarly, http://de.wikipedia.org/Benutzer:DerHexer/Vertrauen lists 483(!) users – some from other projects – that this one user has met in real life

- Thanks for listening
- Questions?
- Wikipedia panel discussion (in German): Wednesday 11:30, Saal 1
- “Hack the Wiki” workshop