A Content-Driven Reputation System for the Wikipedia

Luca de Alfaro

UC Santa Cruz

Joint work with Bo Adler,

UC Santa Cruz

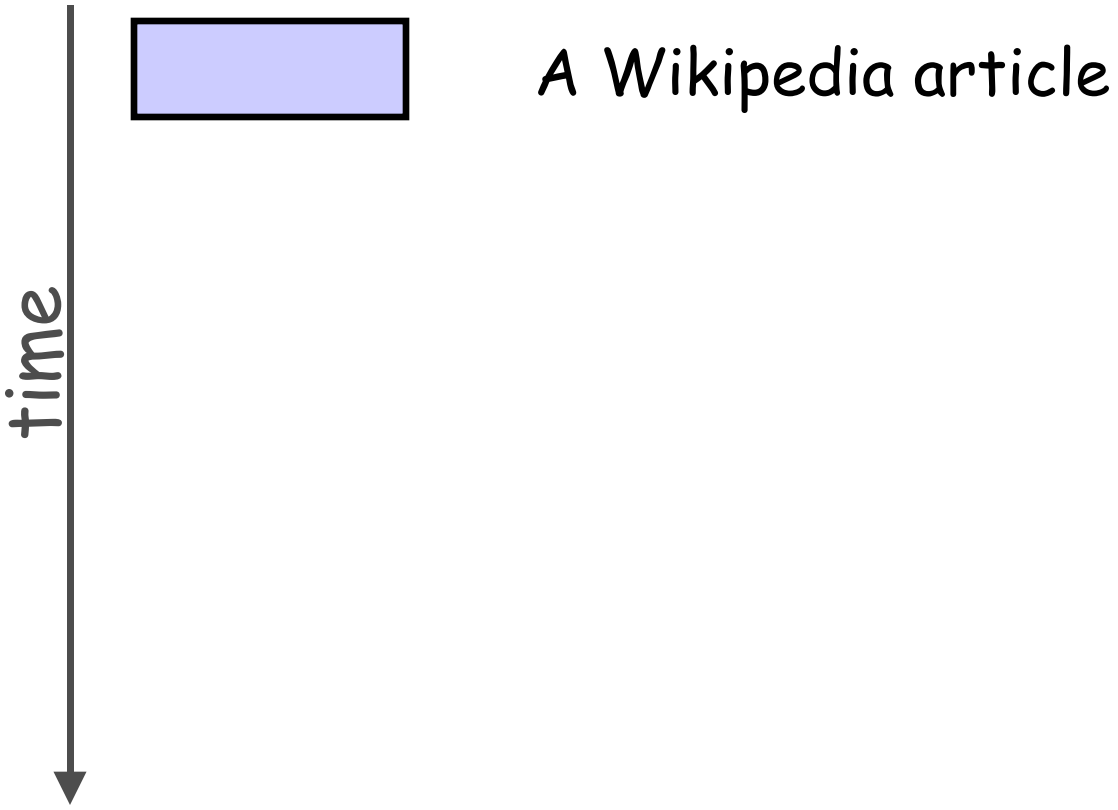
Google, June 2007

Why reputation?

- The Wikipedia is open: almost all pages can be edited by anybody.
- We would like to:
  - **Provide readers with some information on the reliability, or "trust", of the text.** Measuring the reputation of authors is a first step.
  - **Provide an incentive to give lasting contributions.**
  - **Provide a basis for limiting editing to controversial or high-visibility pages.**

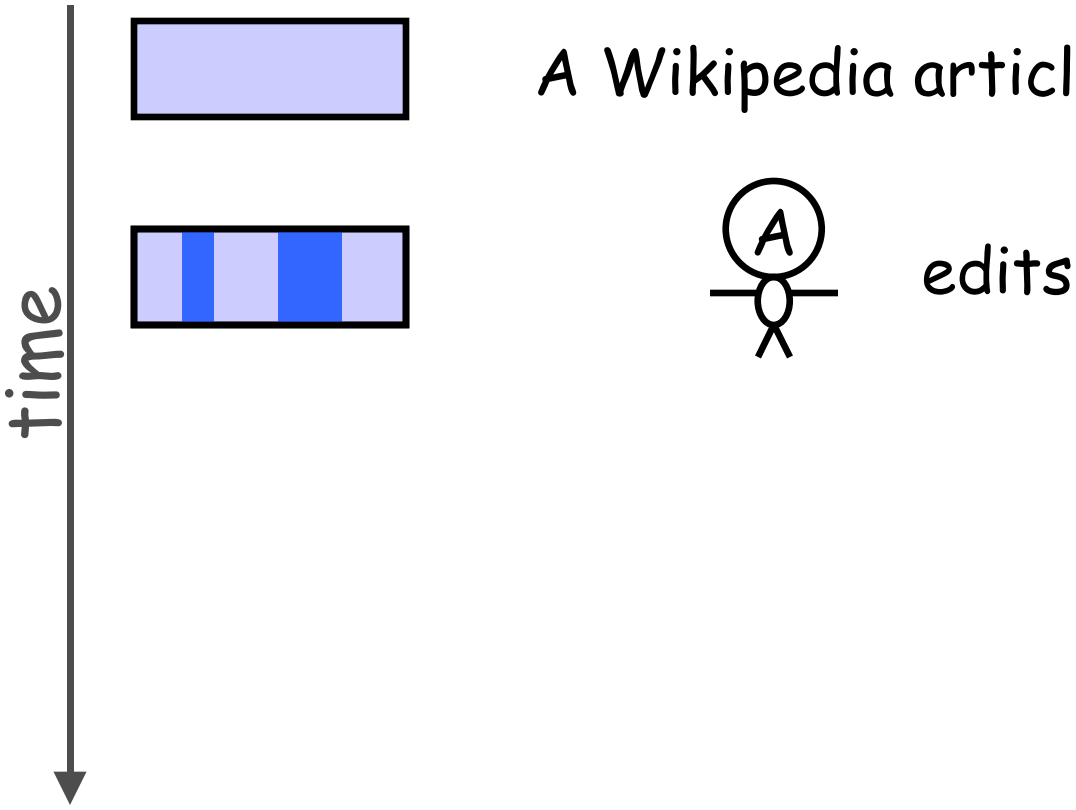
Content-driven reputation

- **Authors of long-lived contributions gain reputation**
- **Authors of reverted contributions lose reputation**



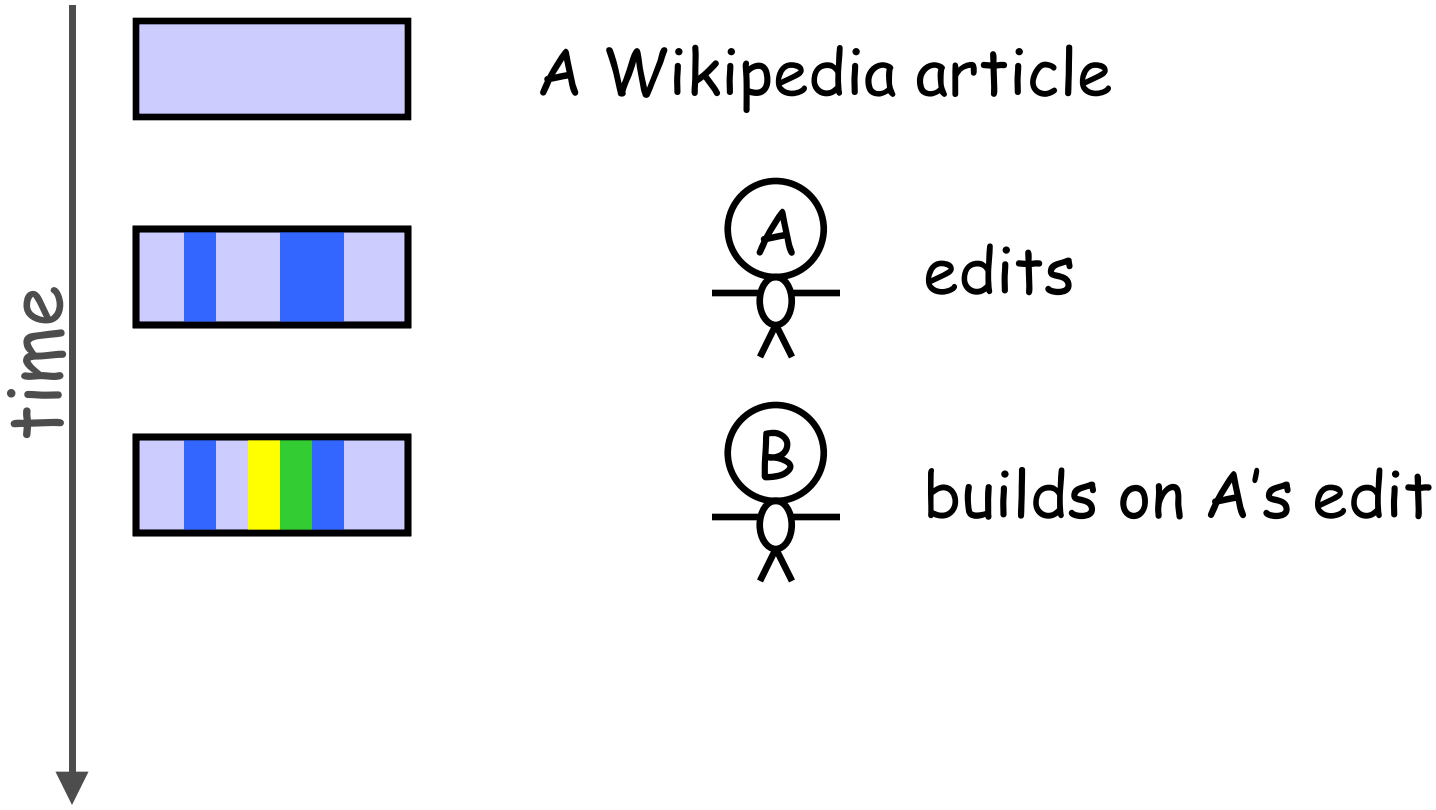
Content-driven reputation

- **Authors of long-lived contributions gain reputation**
- **Authors of reverted contributions lose reputation**



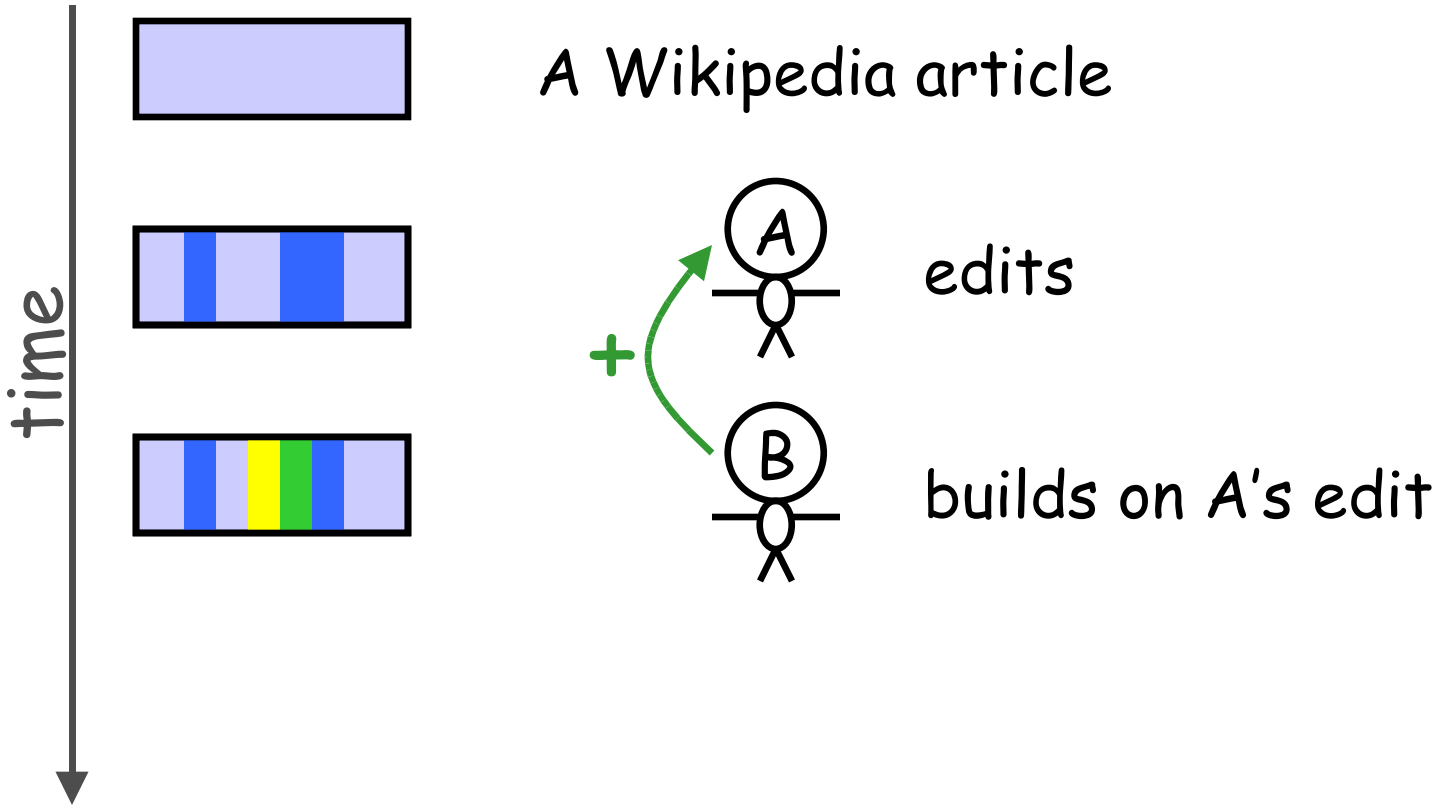
Content-driven reputation

- **Authors of long-lived contributions gain reputation**
- **Authors of reverted contributions lose reputation**



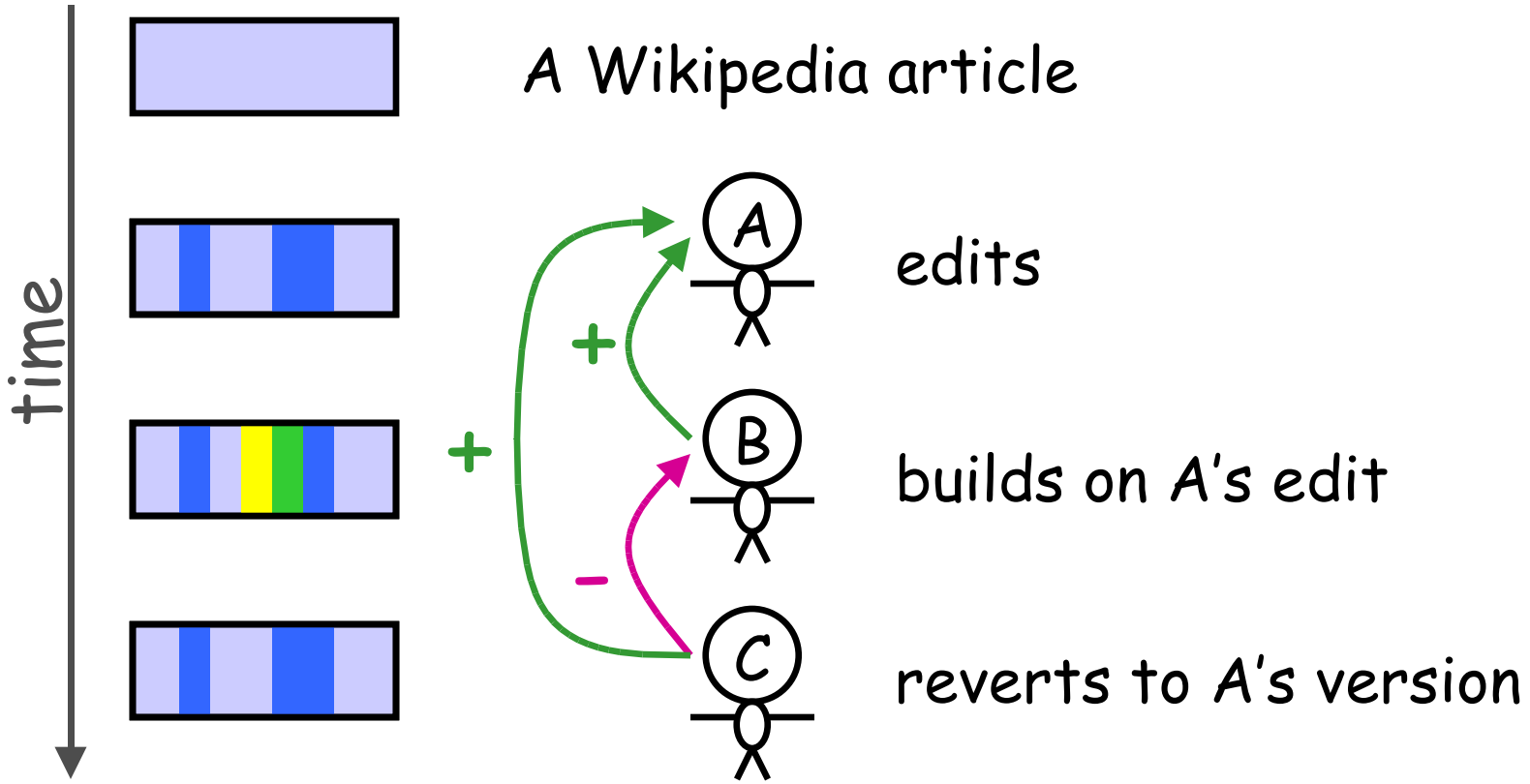
Content-driven reputation

- **Authors of long-lived contributions gain reputation**
- **Authors of reverted contributions lose reputation**



Content-driven reputation

- **Authors of long-lived contributions gain reputation**
- **Authors of reverted contributions lose reputation**



Why content-driven reputation?

Many (most?) reputation systems are based on user-to-user comments (e.g., eBay).

Content-driven reputation is worth investigating:

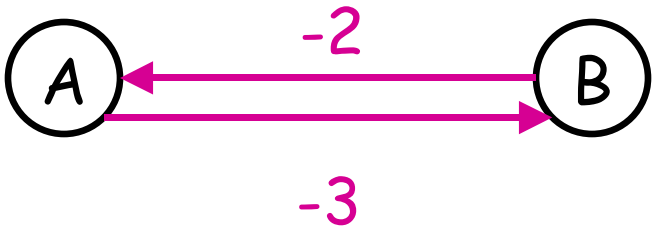
- **Less social stress: makes people feel more welcome.**
  - No need to explicitly rate others, and be rated.
  - Transparent to casual users
  - We would like to avoid displaying reputation directly: only used for access to pages, text trust, ...
- **Avoids reputation wars** (see next slide)
- **We have the data**
  - The whole wikipedia history, many million article versions.

Avoids reputation wars

Wars in user-driven reputation:

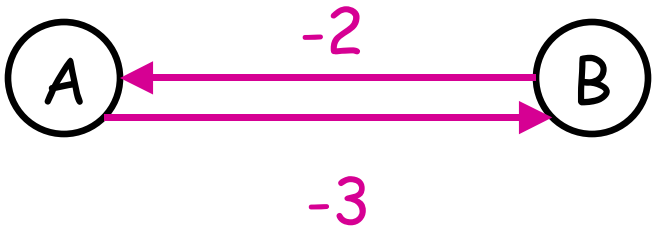
Avoids reputation wars

Wars in user-driven reputation:

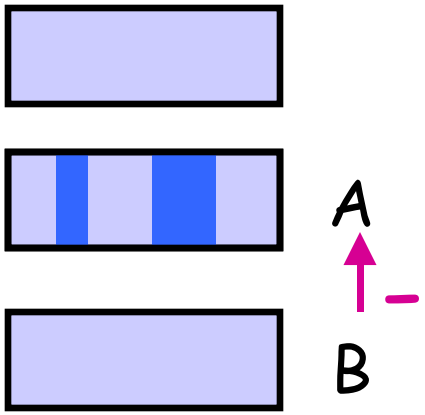


Avoids reputation wars

Wars in user-driven reputation:



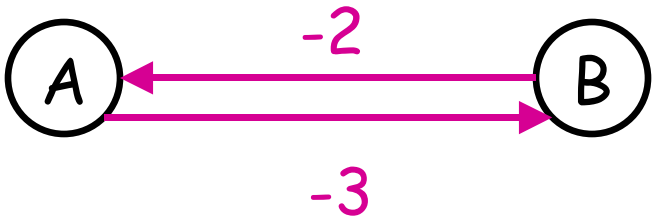
Wars in content-driven reputation:



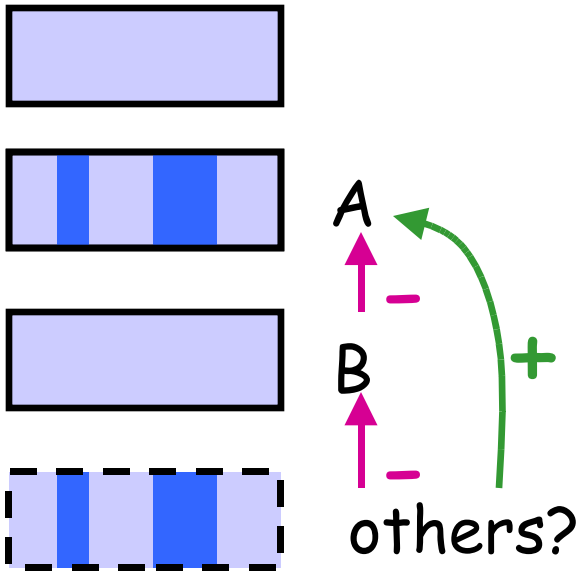
- B can badmouth A by undoing her work

Avoids reputation wars

Wars in user-driven reputation:



Wars in content-driven reputation:

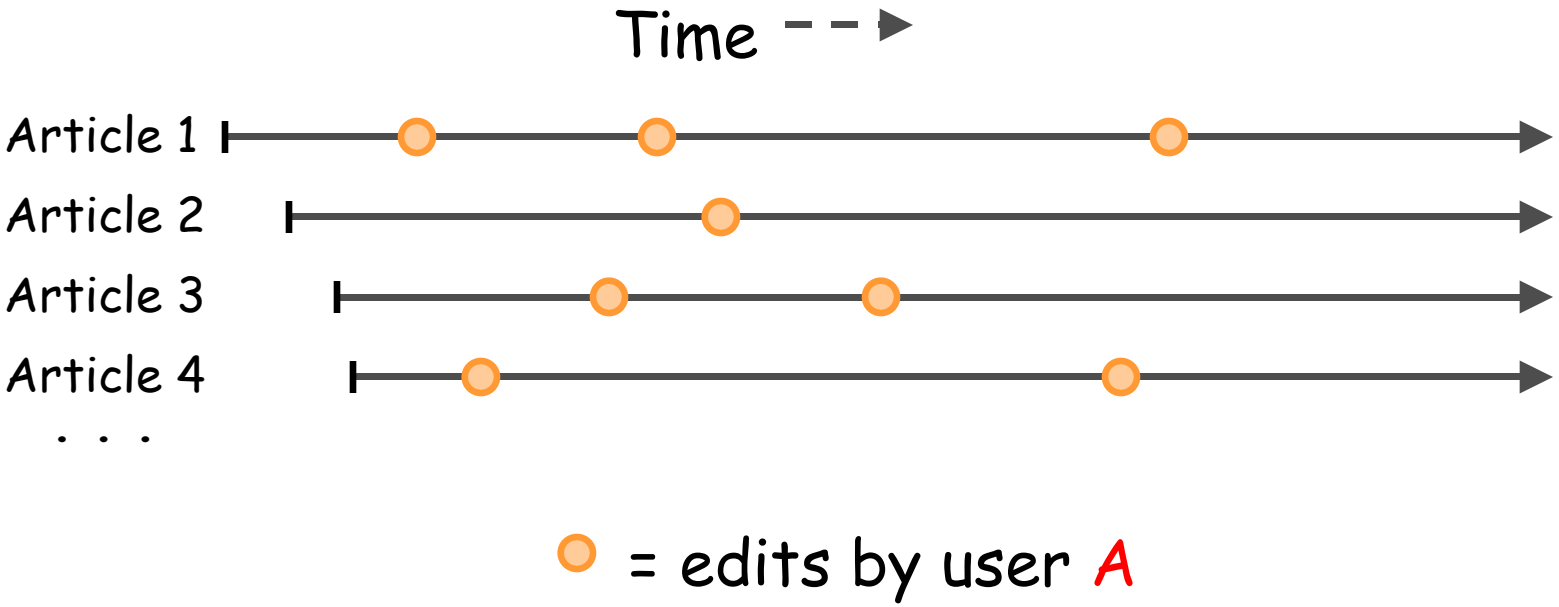


- B can badmouth A by undoing her work
- But this is risky: if others then re-instate A's work, it is B's reputation that suffers.

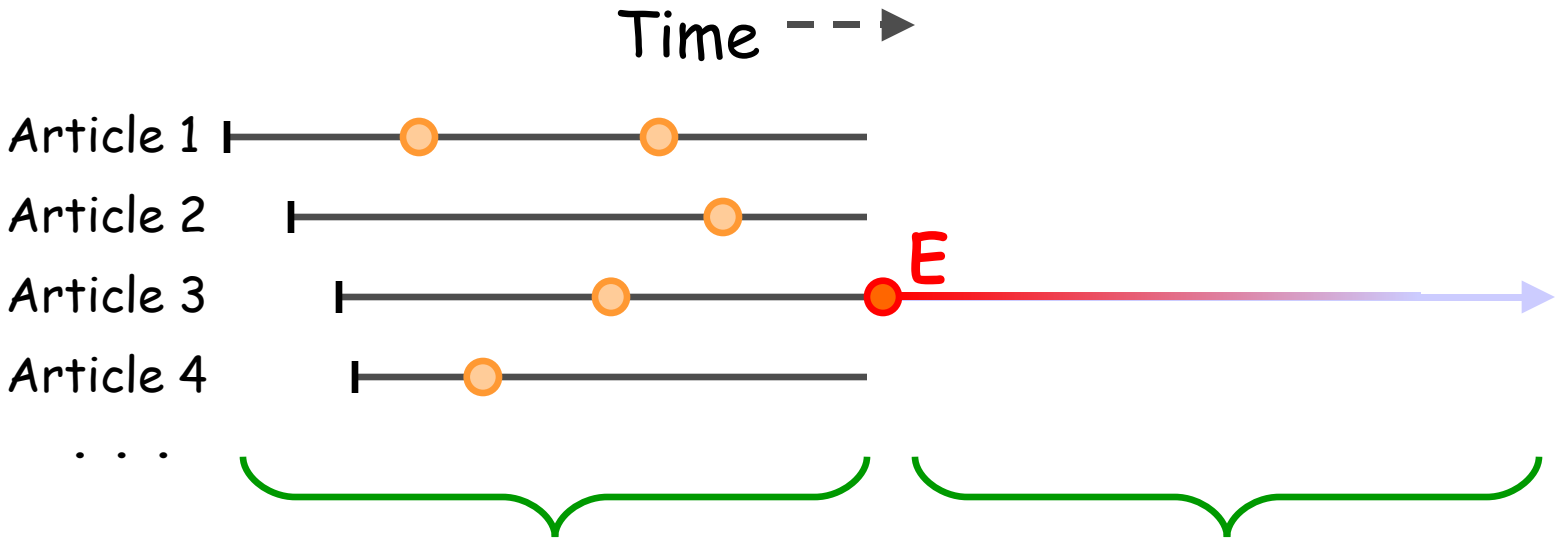
Goals of reputation in Wikipedia

- **Prescriptive:** encourages people to behave in a good way (e.g., Ebay system).
  - We want to encourage lasting contributions.
- **Descriptive:** gives information to users (e.g., Pagerank, Ebay system).
  - Author reputation can be used as a rough guide to the trust in new text/edits.
- **Predictive:** Is reputation a good predictor for future behaviour? Few systems make this claim!
  - **We use this as our evaluation criterion, and we show that our reputation can predict edit quality.**

Validation: Does our reputation have predictive value?



Validation: Does our reputation have predictive value?



The **reputation of author A** at the time of an edit **E** depends on the history **before** the edit.

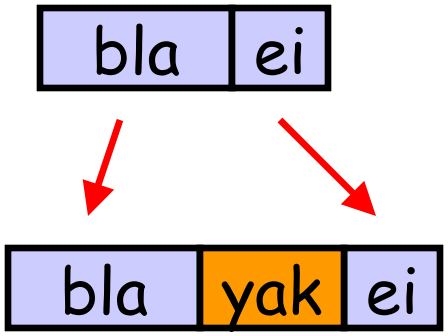
The **longevity** of an edit **E** depends on the history **after** the edit.

Can we show a correlation between author reputation and edit longevity ?

Building a content-driven reputation system

What is a "contribution"?

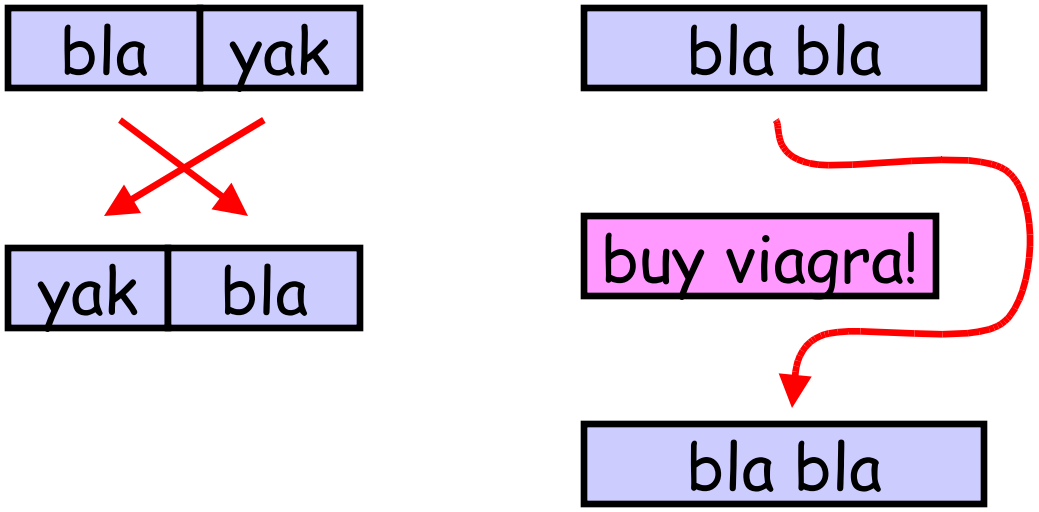
Text



We measure how long the added text survives.

Based on text tracking.

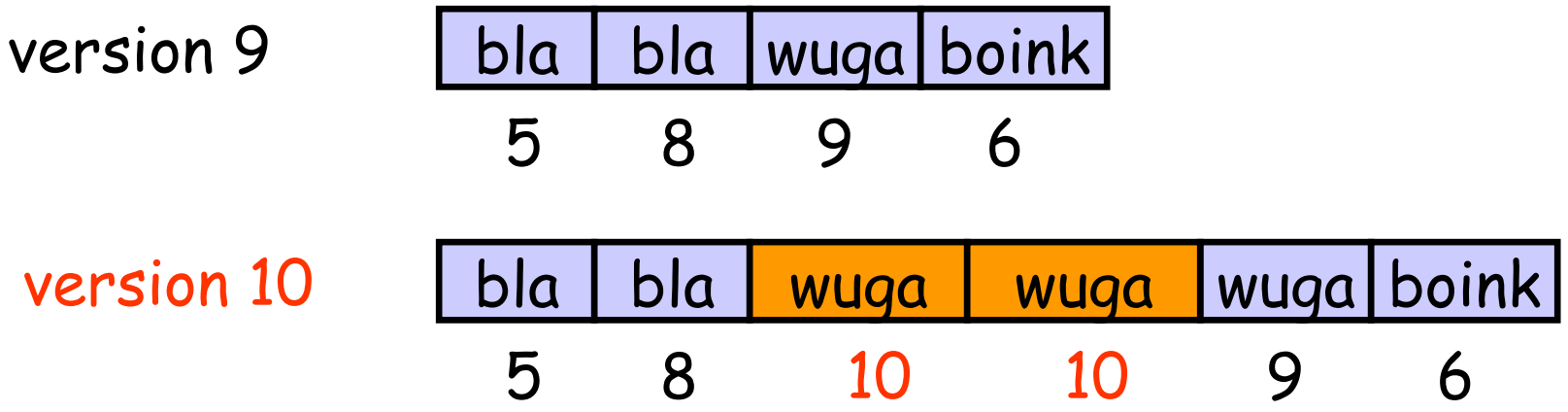
Edit



We measure how long the "edit" (reorganization) survives.

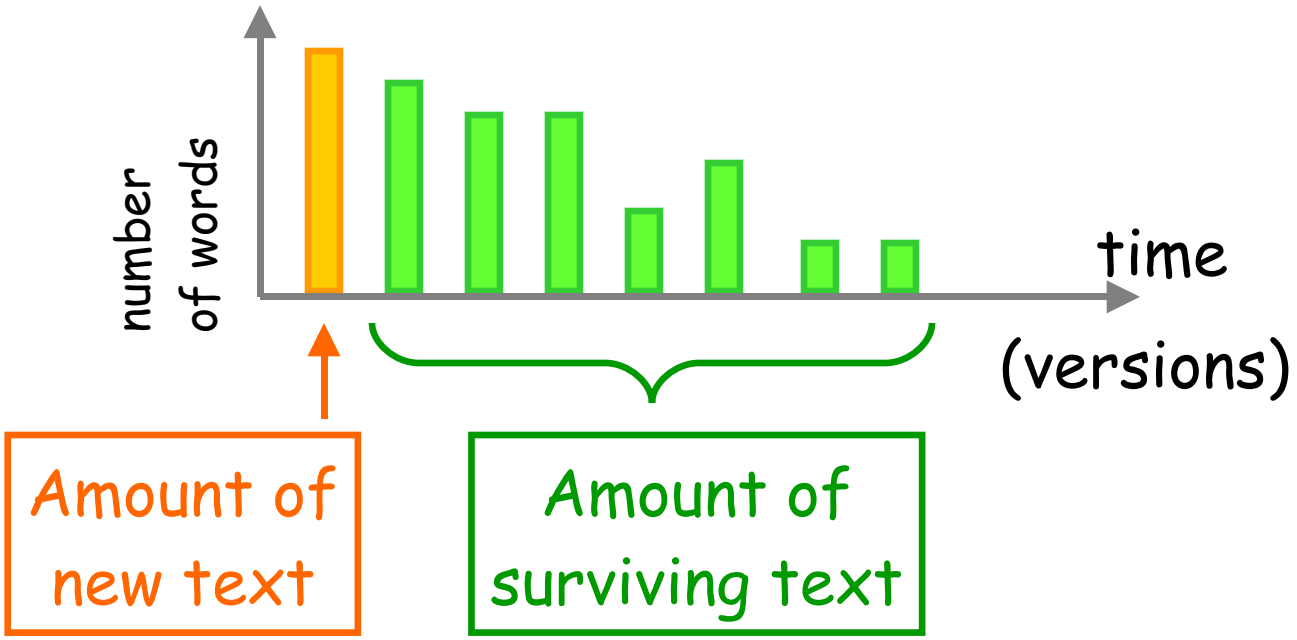
Based on edit distance.

Text



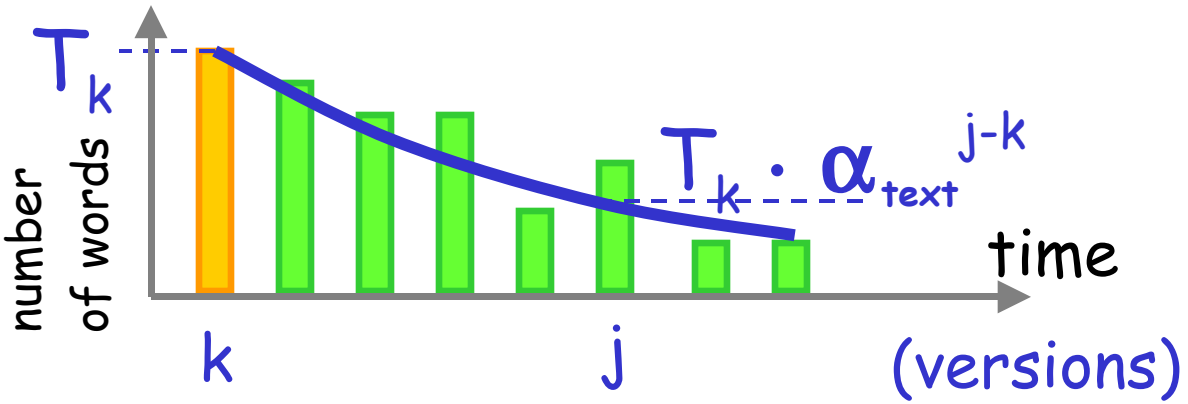
We label each word with the version where it was introduced. This enables us to keep track of how long it lives.

Text



The life of the text introduced at a revision.

Text: Longevity

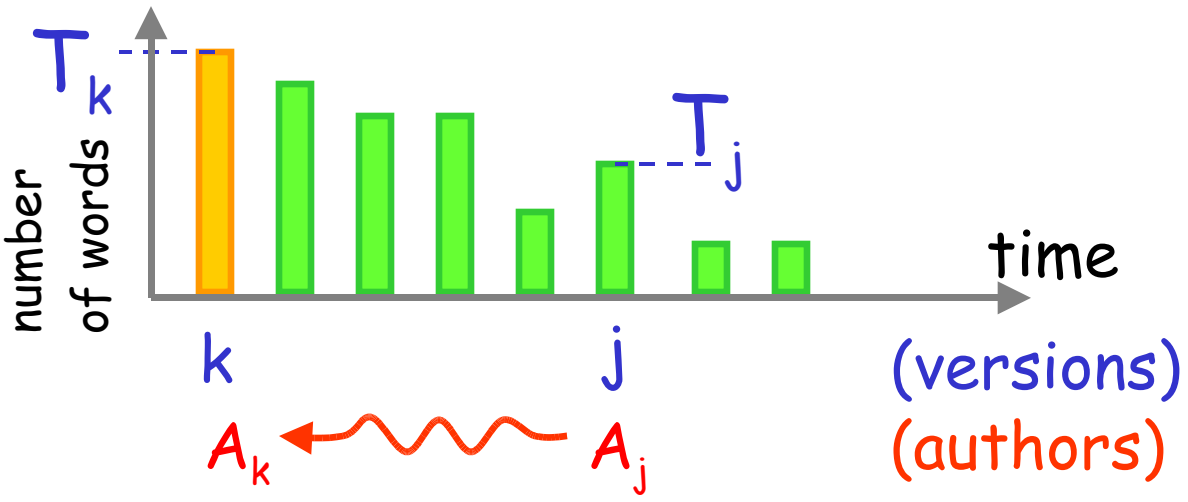


Text Longevity α_{text}:

- We find the $\alpha_{text} \in [0,1]$ that yields the best geometrical approximation for the amount of residual text.
- We call α_{text} the ***text longevity*** of edit k.

  $\alpha_{text} \simeq 1$: long-lived; $\alpha_{text} = 0$: removed immediately.

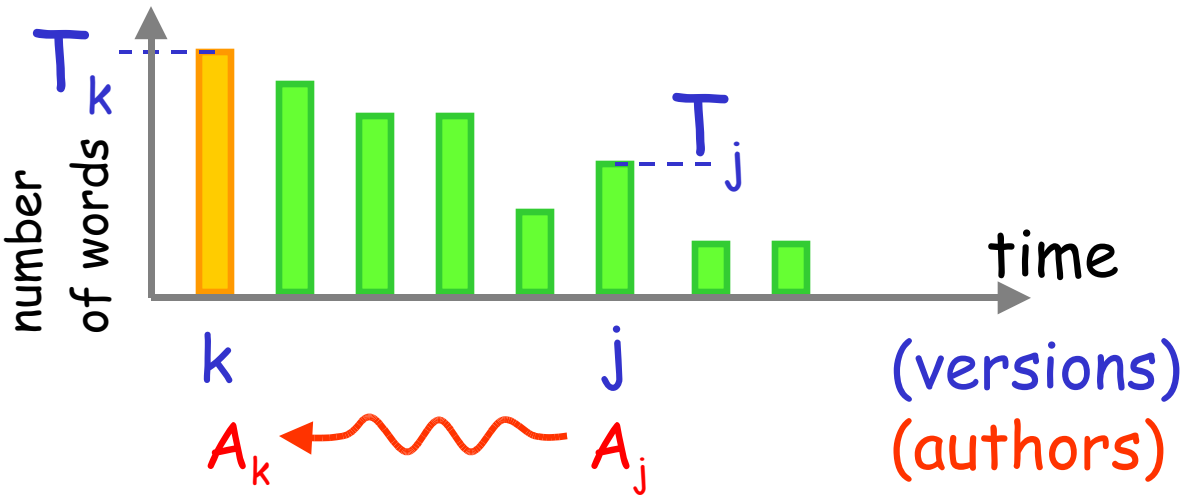
Text: Reputation update



As a consequence of version j, we increment the reputation of A_k by:

$$c_{text} \cdot c_{rep} \cdot \frac{T_j}{T_k} \cdot T_k^{c_{len}} \cdot \log\left(1 + rep(A_j)\right)$$

Text: Reputation update



As a consequence of version j, we increment the reputation of A_k by:

$$c_{text} \cdot c_{rep} \cdot \frac{T_j}{T_k} \cdot T_k^{c_{len}} \cdot \log\left(1 + rep(A_j)\right)$$

Determined via learning; goal: optimize predictive value.

Measuring text life: Issues

Duplication attack:

Version

9

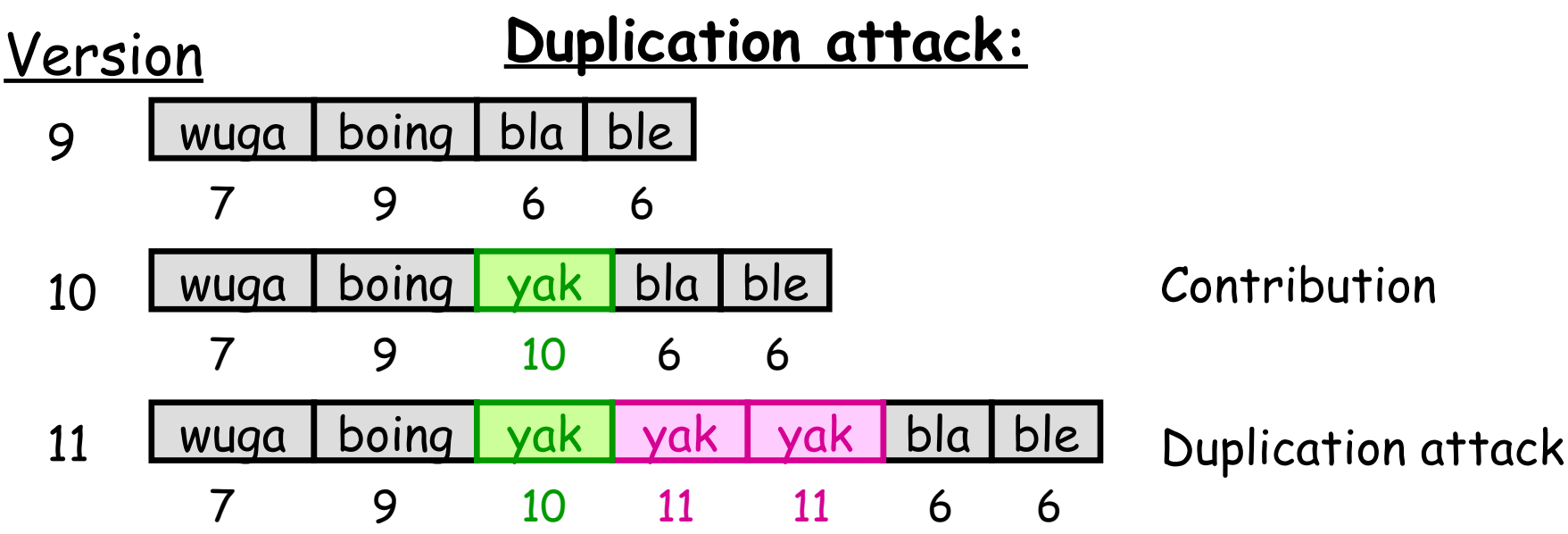
7 9 6 6

10

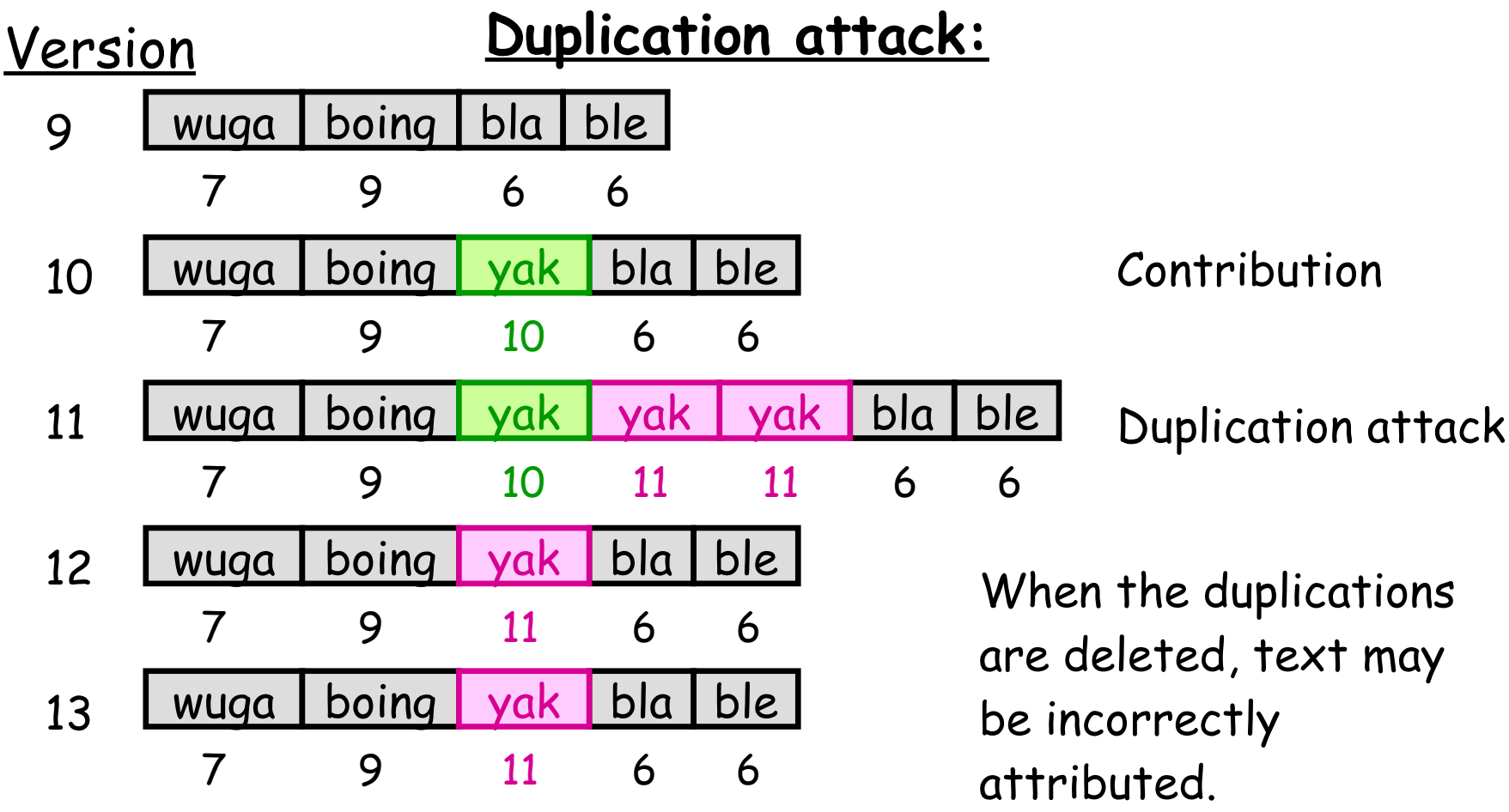
7 9 10 6 6

Contribution

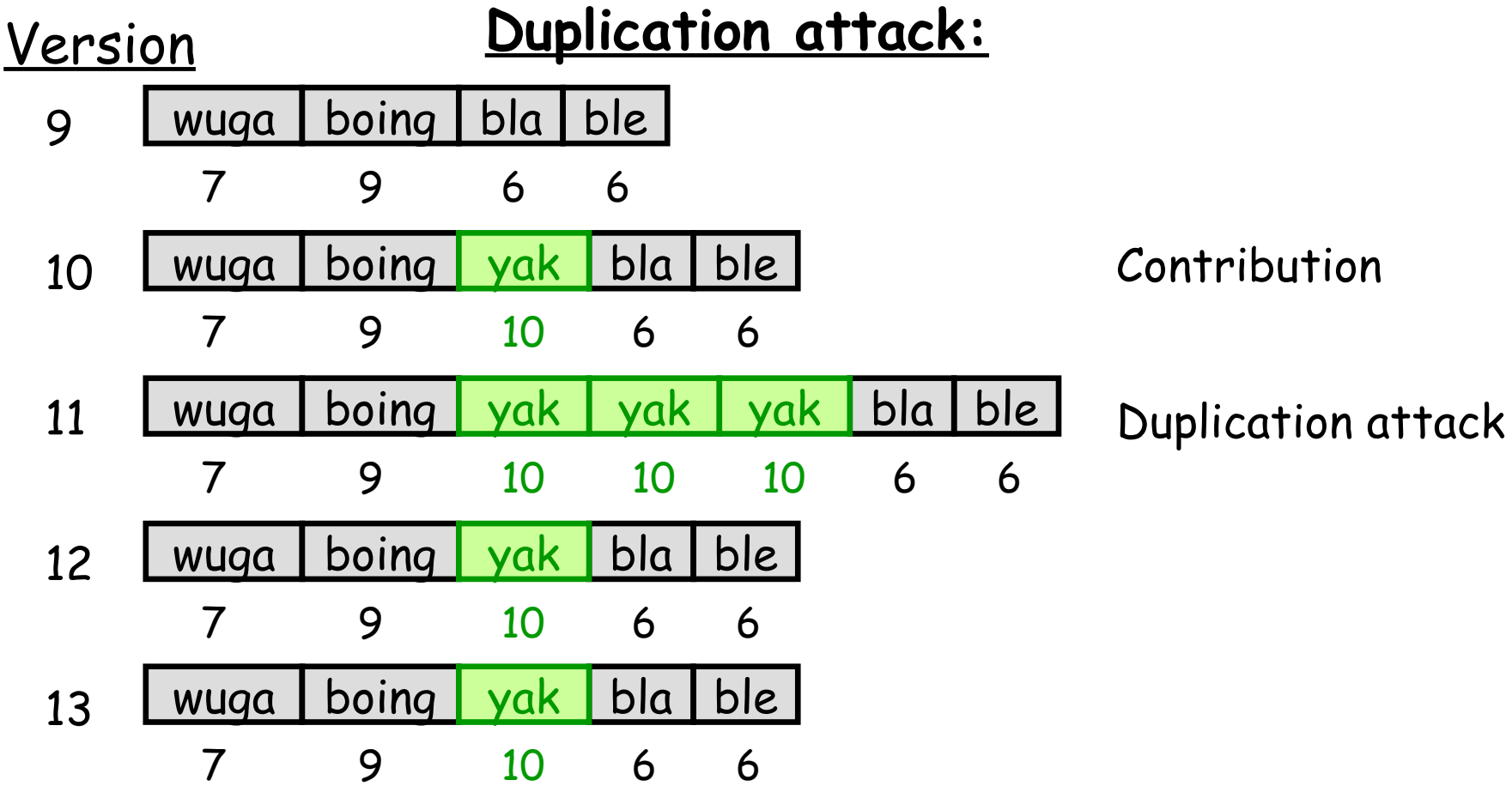
Measuring text life: Issues



Measuring text life: Issues



Measuring text life: Issues

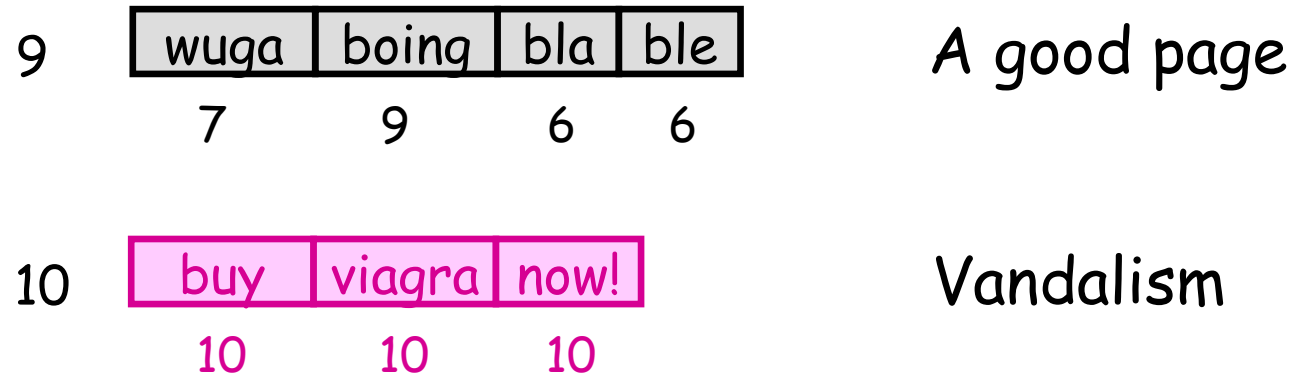


Solution: text can be matched multiple times.

Measuring text life: Issues

Dealing with reversions

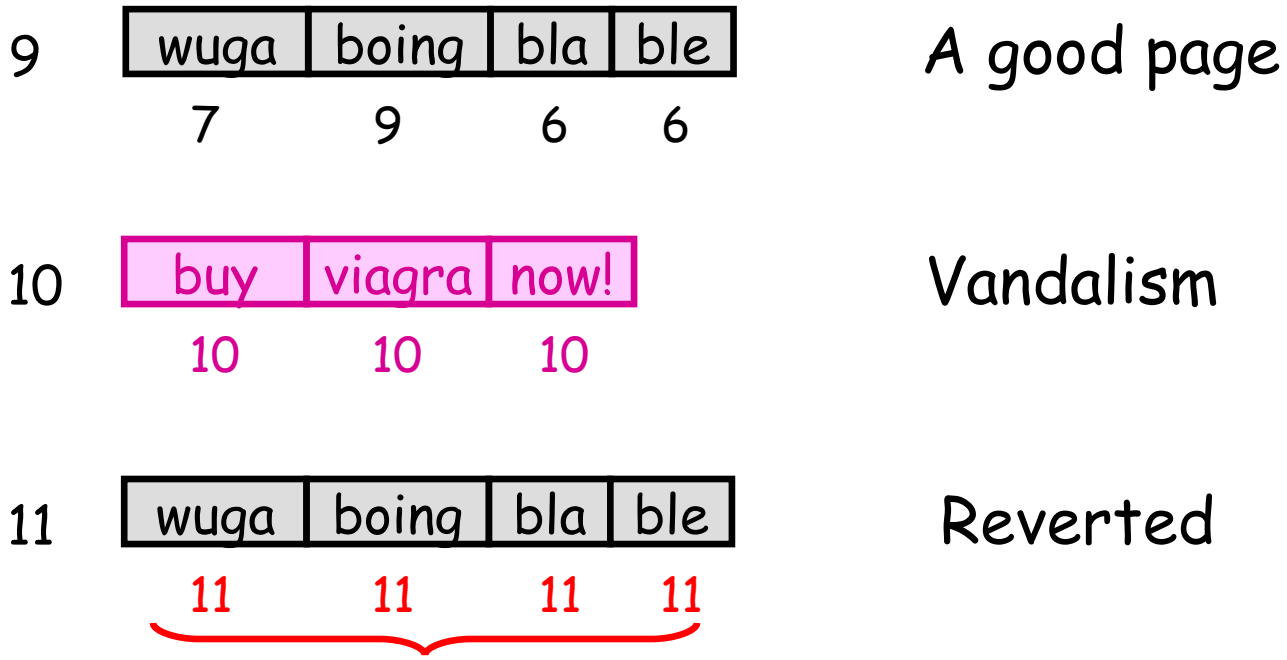
Version



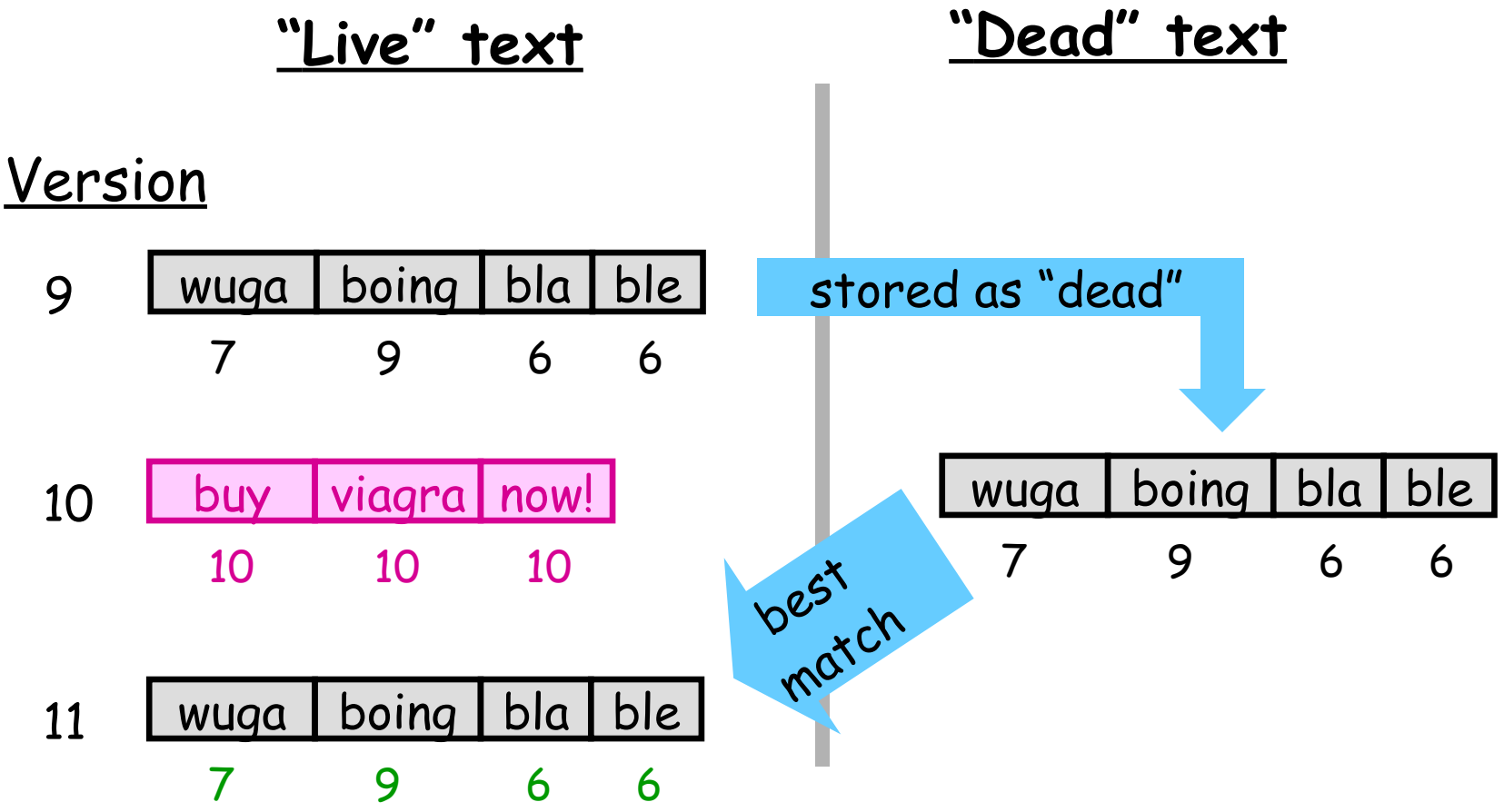
Measuring text life: Issues

Dealing with reversions

Version

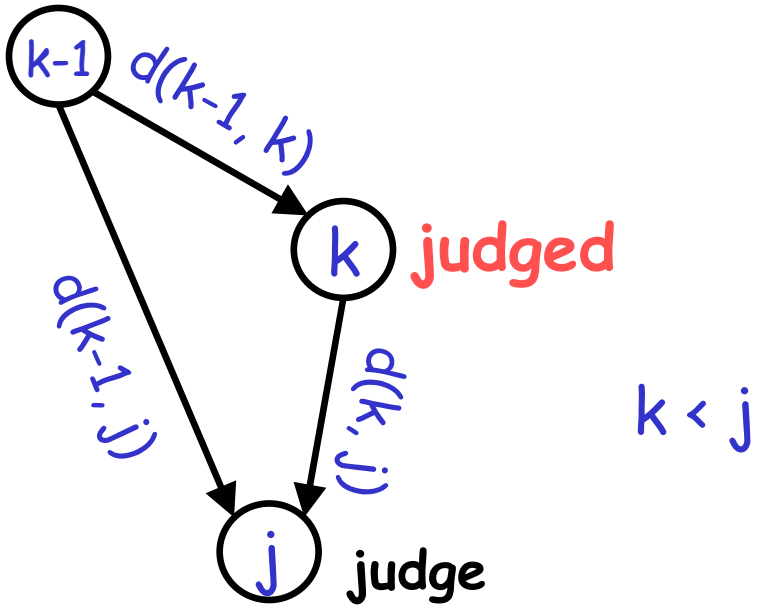


Measuring surviving text



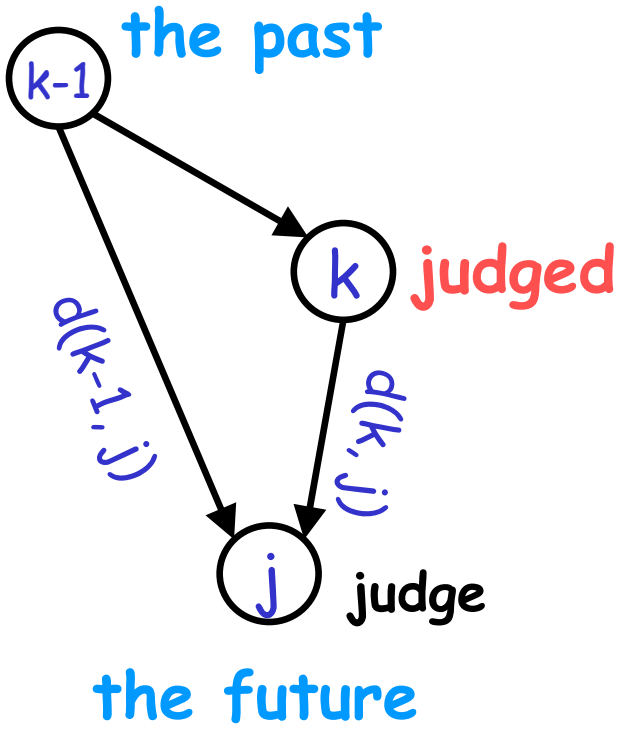
We track authorship of deleted text, and we match the text of new versions both with live and with dead text.

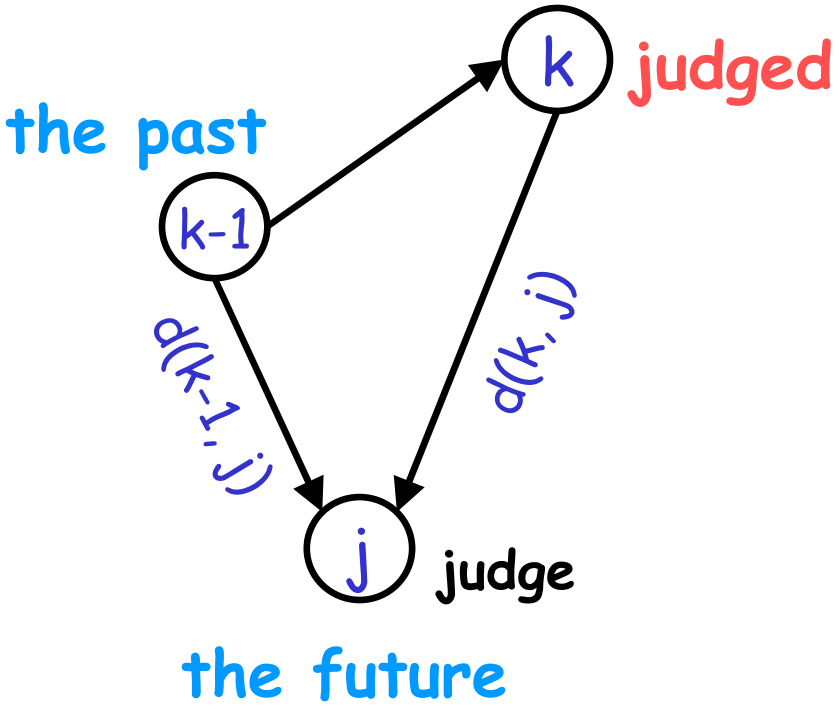
Edit



We compute the edit distance between versions k-1, k, and j, with k < j (see paper for details on the distance).

Edit: good or bad?





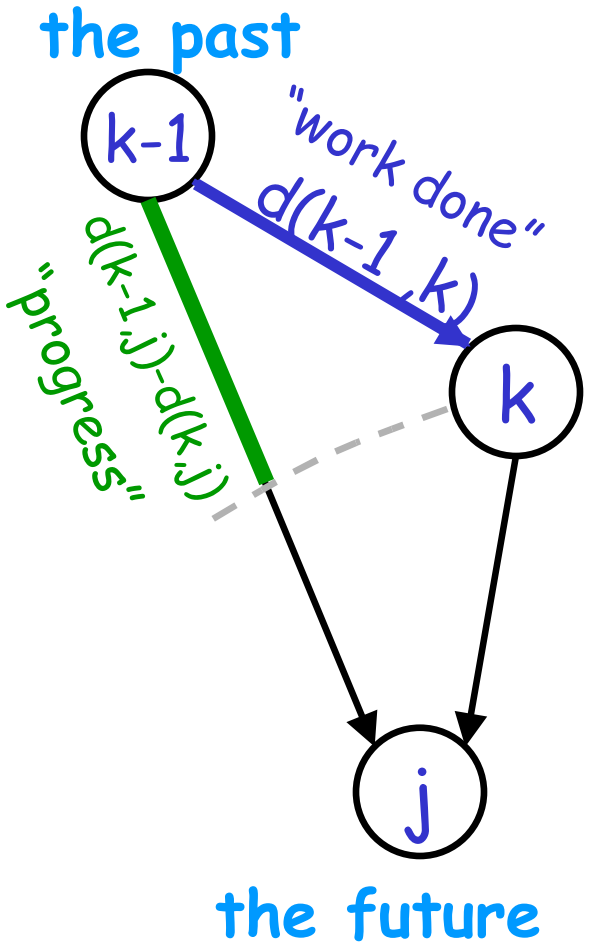
k is **good**: d(k-1, j) > d(k, j)

"k went towards the future"

k is **bad**: d(k-1, j) < d(k, j)

"k went against the future"

Edit: Longevity



Edit Longevity:

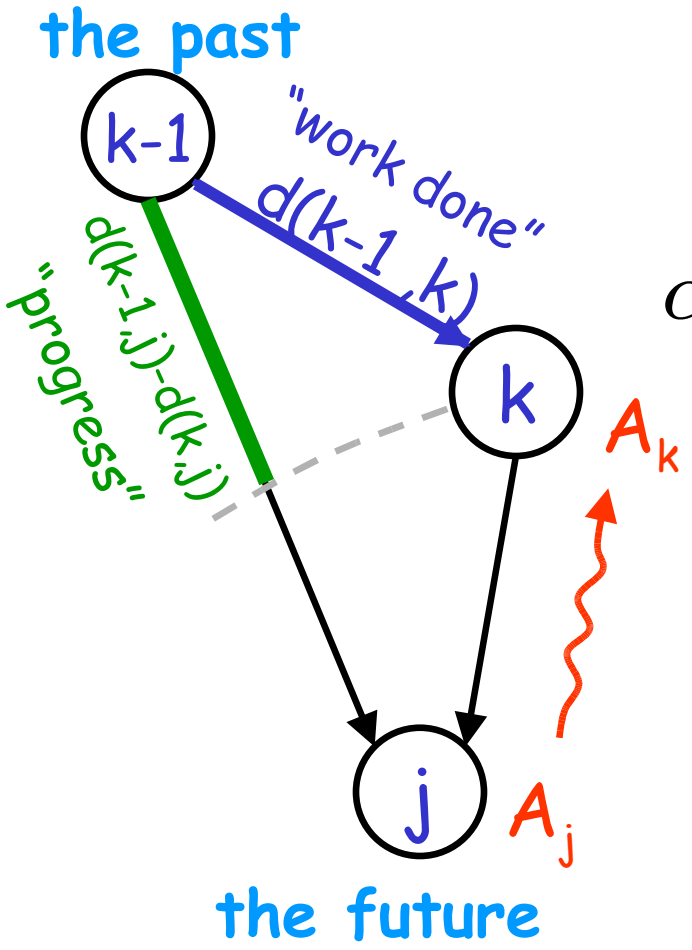
$$\alpha_{edit} = \frac{d(k-1,j) - d(k,j)}{d(k-1,k)}$$

The fraction of change that is in the same direction of the future.

- $\alpha_{edit} \simeq 1$: k is a good edit
- $\alpha_{edit} \simeq -1$: k is reverted

Corollary: we can detect reversions automatically!!

Edit: Updating reputation



$$\alpha_{edit} = \frac{d(k-1,j) - d(k,j)}{d(k-1,k)}$$

Edit Longevity:

Reputation update:

The reputation of A_k

- increases if $\alpha_{edit} > 0$,
- decreases if $\alpha_{edit} < 0$.

(see paper for details)

Computing edit distance

We compare an old version u and a new version v, and we compute:

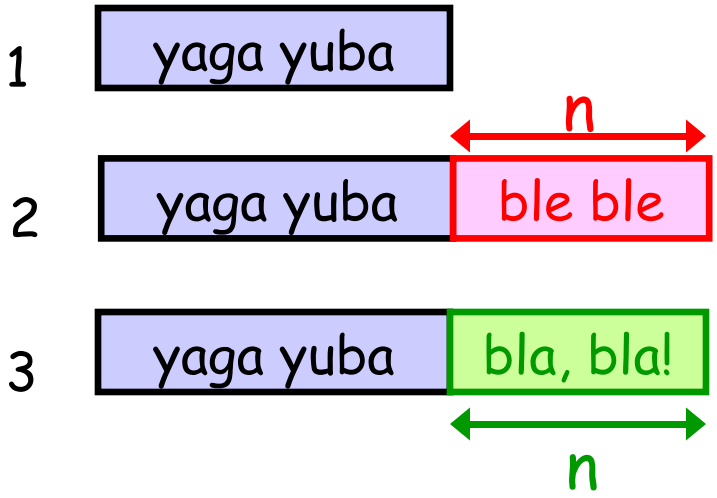
- I_{tot}: Total inserted text
- D_{tot}: Total deleted text
- M_{tot}: Total amount of relative exchange of position (a measure of text reordering).

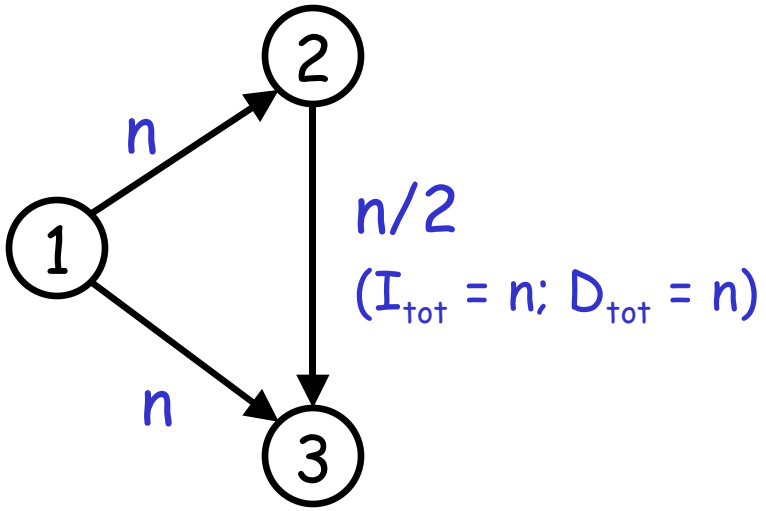
$$d(u,v) = \max(I_{tot}, D_{tot}) - \frac{1}{2}\min(I_{tot}, D_{tot}) + M_{tot}$$

Computing edit distance

$$d(u,v) = \max(I_{tot}, D_{tot}) - \tfrac{1}{2}\min(I_{tot}, D_{tot}) + M_{tot}$$

Dealing with rewrites:





For the author of version 2: $\alpha = \dfrac{n - n/2}{n} = \dfrac{1}{2}$

Text matching: Principles

x y z a b c d e a b c b c t u r

x y w a b c g h a b c d e x y z

After lots of experiments, we found that the best is:

- Match in greedy fashion longest common subsequences first.
- But, give some preference to subsequences occurring in similar relative position in the overall text.

Text matching: Principles

x y z a b c d e a b c b c t u r

x y w a b c g h a b c d e x y z

After lots of experiments, we found that the best is:

- Match in greedy fashion longest common subsequences first.
- But, give some preference to subsequences occurring in similar relative position in the overall text.

Text matching: Principles

x y z a b c d e a b c b c t u r

x y w a b c g h a b c d e x y z

After lots of experiments, we found that the best is:

- Match in greedy fashion longest common subsequences first.
- But, give some preference to subsequences occurring in similar relative position in the overall text.

Text matching: Principles

No:
relative position
is too different

After lots of experiments, we found that the best is:

- Match in greedy fashion longest common subsequences first.
- But, give some preference to subsequences occurring in similar relative position in the overall text.

Text matching: Principles

This is preferred.

After lots of experiments, we found that the best is:

- Match in greedy fashion longest common subsequences first.
- But, give some preference to subsequences occurring in similar relative position in the overall text.

Learning to maximize predictive value

We learned the coefficients in our reputation update formulas, with the goal of maximizing the predictive value of reputation (recall × precision, ...).

Training Wikipedia:

- Italian till Oct 05 154,261 pages, 714,280 versions

Evaluation Wikipedias:

- English till Feb 07 1,988,627 pages, 40,455,416 versions
- French till Feb 07 452,577 pages, 5,643,636 versions
- Italian till May 07 301,584 pages, 3,129,453 versions

Results: English Wikipedia, in detail

% of edits below a given longevity

log (1 + reputation)

Bin	%_data	l<0.8	l<0.4	l<0.0	l<-0.4	l<-0.8
0	16.922	93.11	91.65	89.15	83.76	73.53
1	1.191	77.24	69.83	65.60	61.11	56.00
2	1.335	69.53	57.08	49.79	45.71	41.25
3	1.627	38.00	28.61	20.23	16.16	13.62
4	2.780	32.84	22.31	13.32	9.57	8.04
5	4.408	41.70	15.76	5.90	3.80	2.57
6	6.698	29.40	16.74	7.54	4.35	3.12
7	8.281	32.04	15.16	5.44	2.25	1.40
8	12.233	34.06	16.64	6.78	3.79	2.73
9	44.524	32.55	15.51	5.05	1.88	1.14

Results: English Wikipedia, in detail

% of edits below a given longevity

log (1 + reputation)

low rep

Bin	% data	l<0.8	l<0.4	l<0.0	l<-0.4	l<-0.8
0	16.922	93.11	91.65	89.15	83.76	73.53
1	1.191	77.24	69.83	65.60	61.11	56.00
2	1.335	69.53	57.08	49.79	45.71	41.25
3	1.627	38.00	28.61	20.23	16.16	13.62
4	2.780	32.84	22.31	13.32	9.57	8.04
5	4.408	41.70	15.76	5.90	3.80	2.57
6	6.698	29.40	16.74	7.54	4.35	3.12
7	8.281	32.04	15.16	5.44	2.25	1.40
8	12.233	34.06	16.64	6.78	3.79	2.73
9	44.524	32.55	15.51	5.05	1.88	1.14

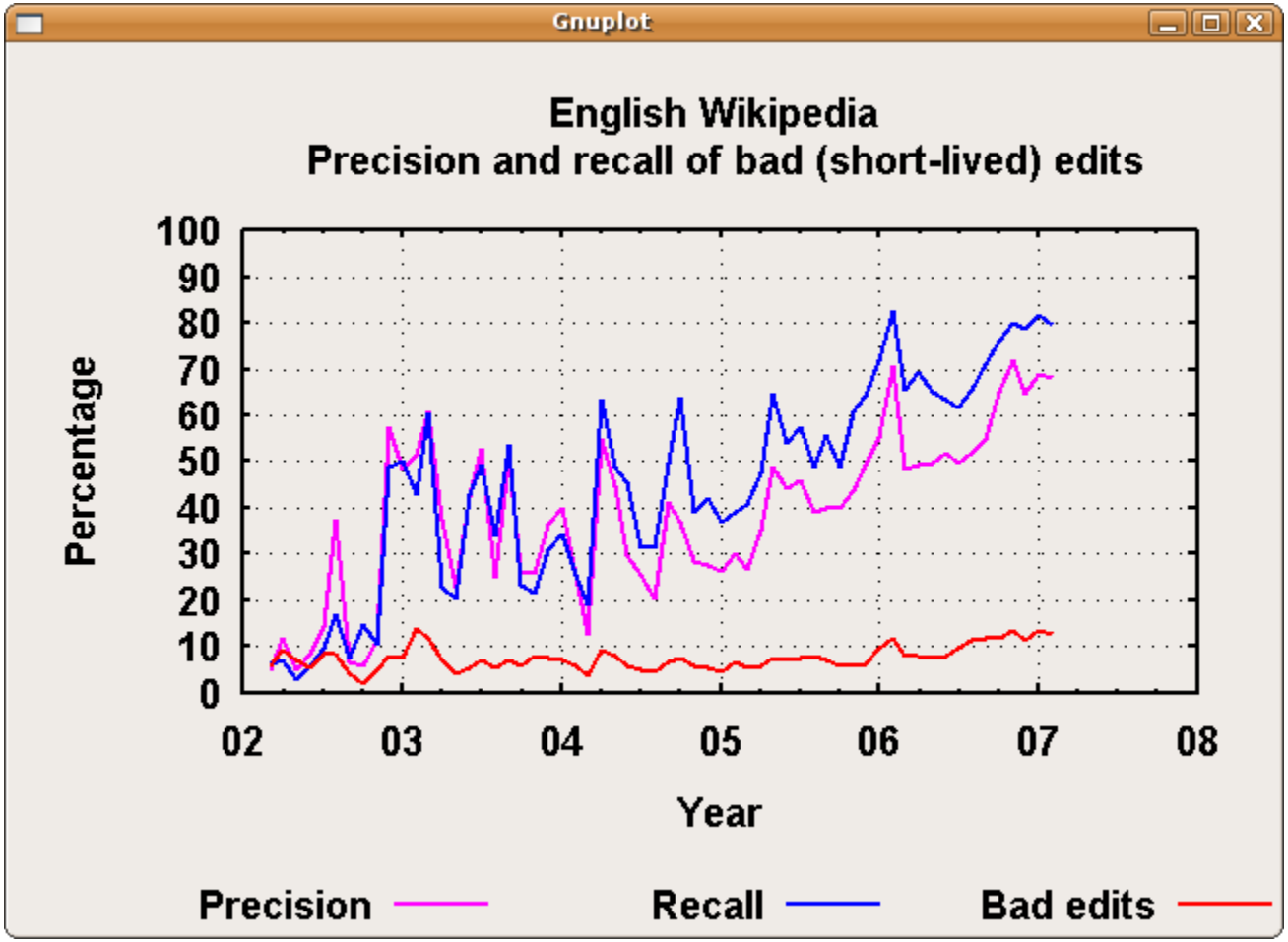
Short-Lived

Predictive power of low reputation

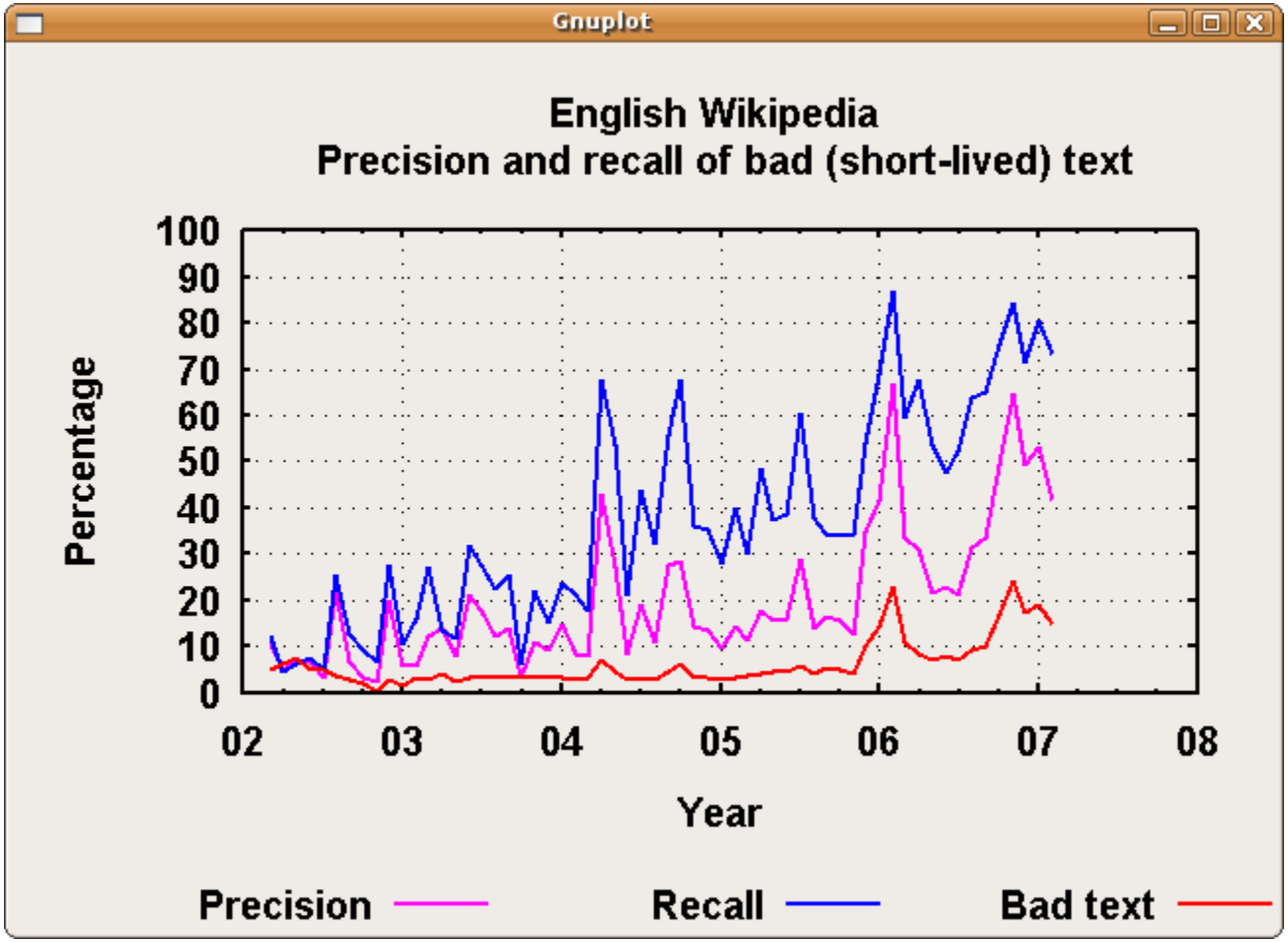
Low-reputation: Lower 20% of range

	Short-lived edits		Short-lived text	
	$\alpha_{edit} \leq -0.8$ (almost entirely undone)		$\alpha_{text} \leq 0.2$ (less than 20% survives each revision)	
Wikipedia	**Precision**	**Recall**	**Precision**	**Recall**
English	74.2 %	84.5 %	62.0 %	83.0 %
French	38.4 %	41.2 %	17.3 %	48.1 %
Italian	21.7 %	30.4 %	6.2 %	21.4 %

Predicting Short-Lived Edits



Predicting Short-Lived Text



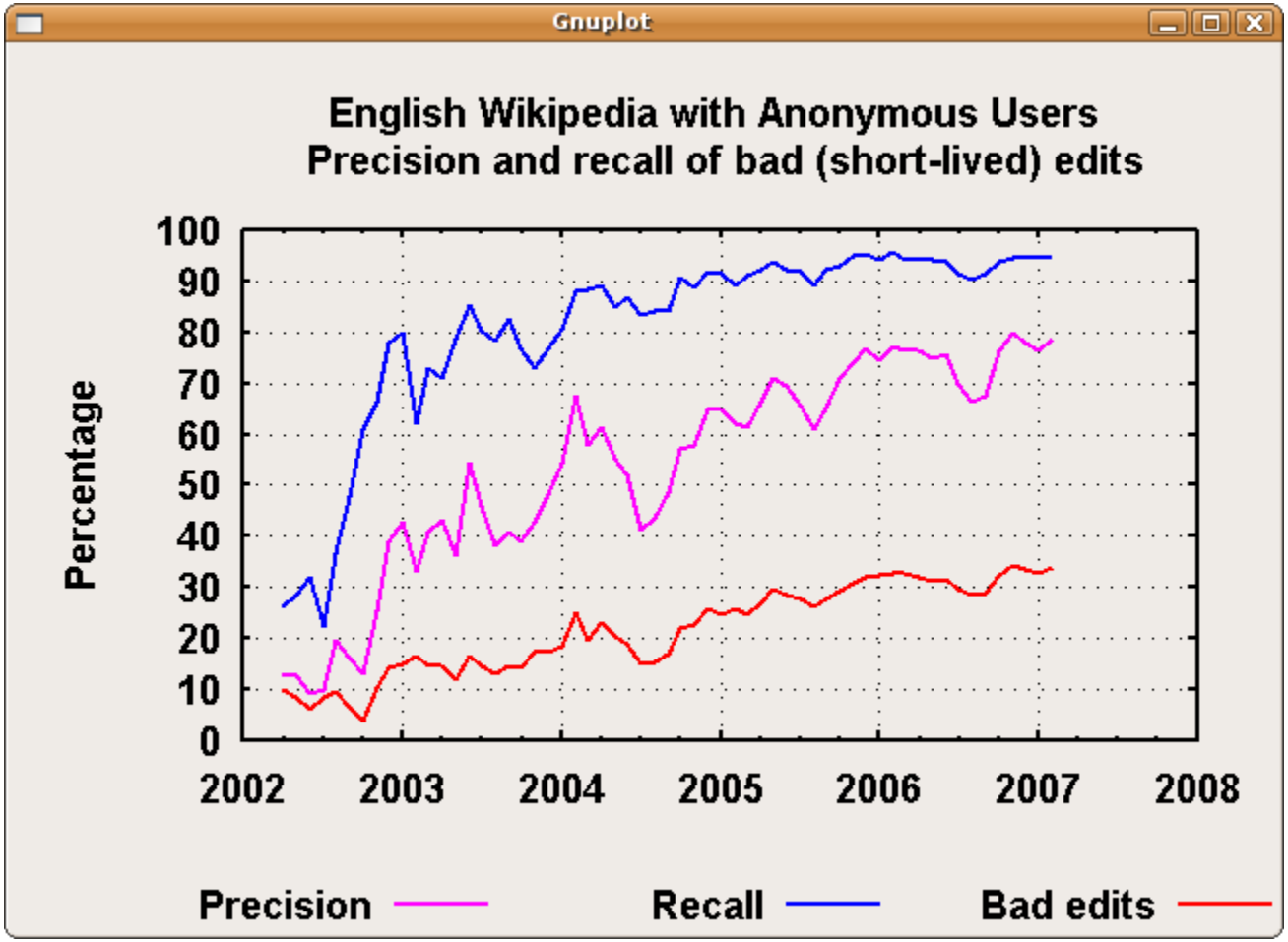
Dealing with Anonymous Authors

- **Registered authors:** register an ID (usually a pseudonym).
- **Anonymous authors:** edit without registering.
  - We assign the lowest reputation to anonymous authors; otherwise, it would be an incentive!

- **Our prediction figures are much better if we include anonymous authors in the statistics:**
  - Anonymous authors are responsible for most of the spam.
  - They do our algorithm a favor: by remaining anonymous, they explicitly signal their non-commitment to the Wikipedia.

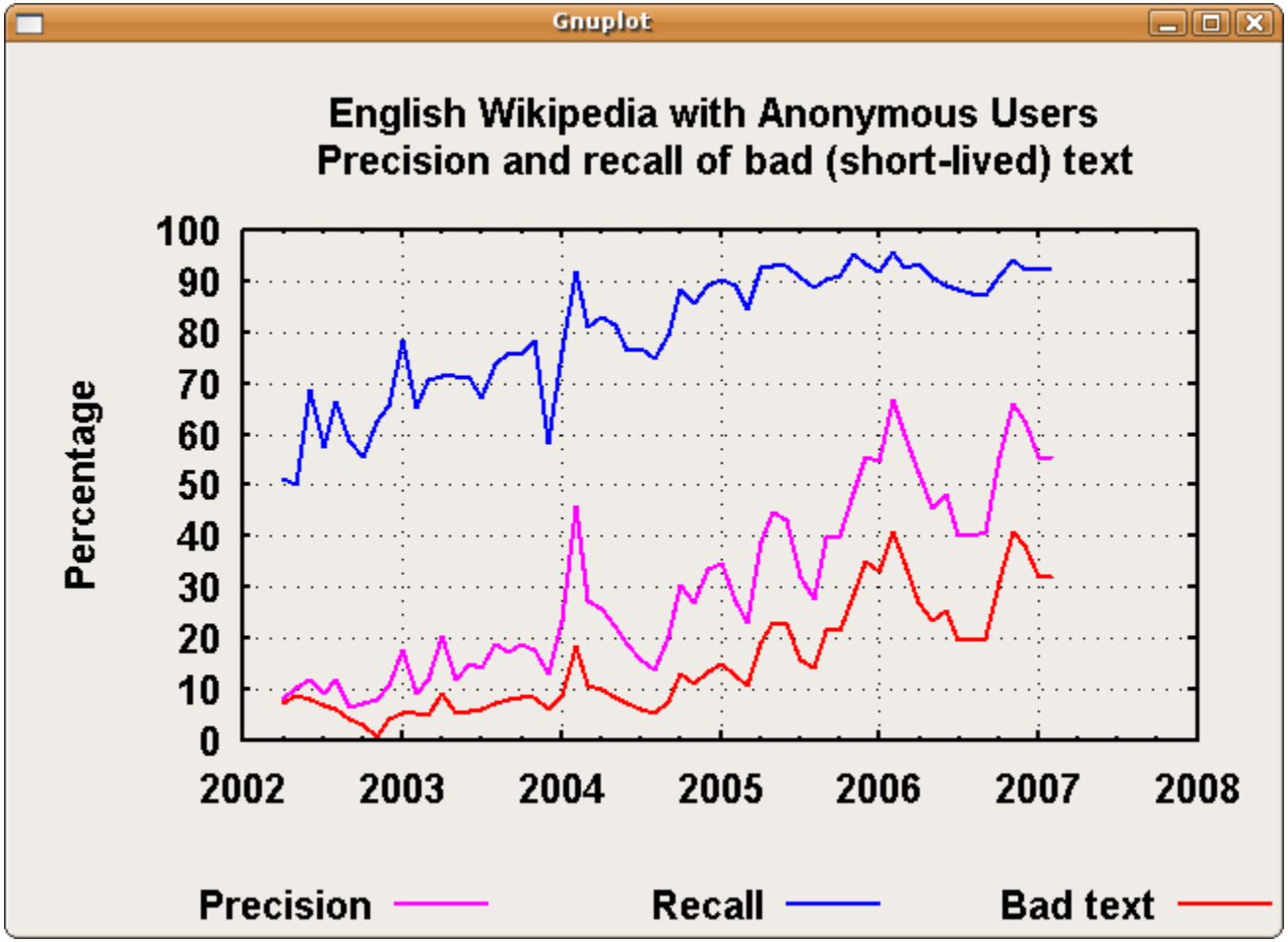
Predicting Short-Lived Edits

Including anonymous authors:



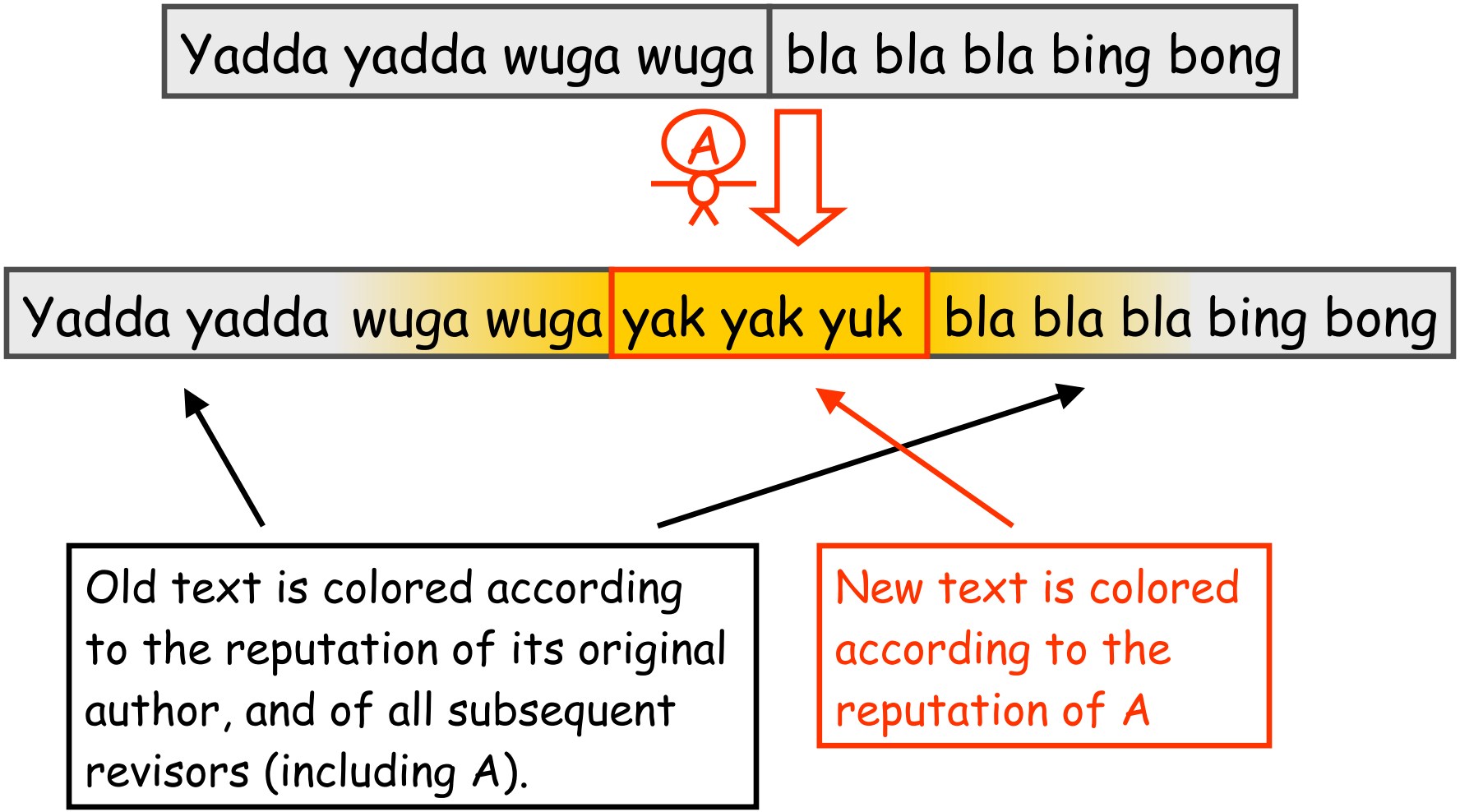
Predicting Short-Lived Text

Including anonymous authors:



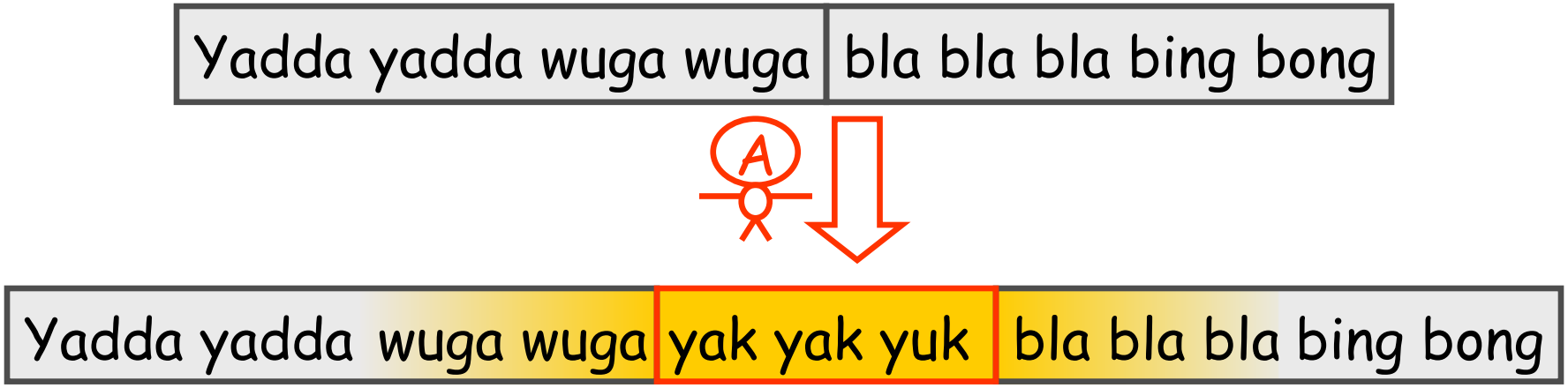
Work in progress:

Author reputation and text trust



Work in progress:

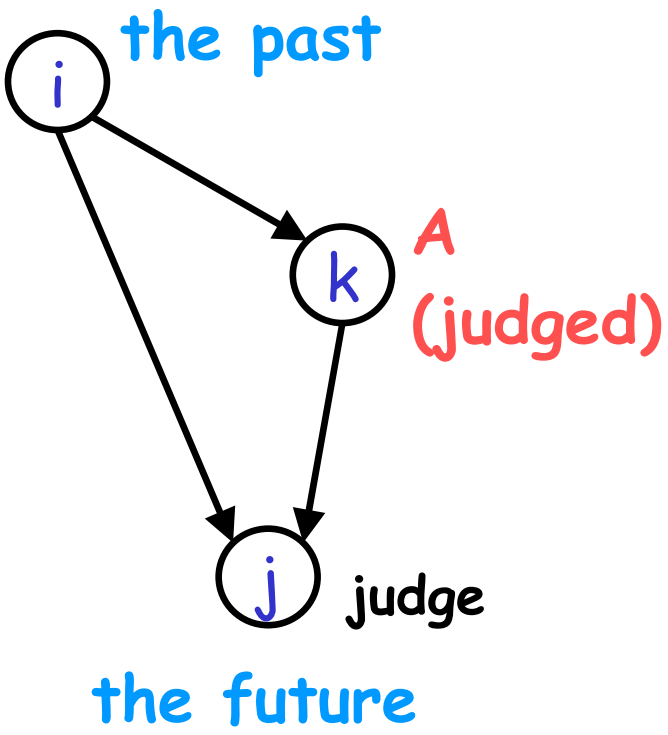
Author reputation and text trust



- If we color text fresh from low-reputation authors, we color over 90% of text which will be short-lived!
- We can spot which text won't make it!
- Text trust is still work in progress...

Work in progress:

Making reputation robust to tampering



Idea:

- To modify the reputation of the author A of k, consider only versions i, j, by authors of equal or greater reputation than A.
- Low-reputation sock puppets are no longer useful.
- Similarly for text.
- How well does this work??

Work in progress:
What's best: generic or topic-specific reputation?

What is a better predictor of the quality of an edit in Mathematics:

- the overall reputation of an author,
- or the reputation computed only wrt. Mathematics pages?

No need to philosophize: we can measure this!

Gillian Smith is working on this at UCSC.

What's your bet?

Social interaction on the Wikipedia

- We can automatically detect reversions, edit wars, and other phenomena!
- This enables all sorts of sociological studies of the Wikipedia, and of how people collaborate.
- We are just starting to look into the data...

Some lighter fare…

What are the articles of the Italian Wikipedia where the biggest edit wars occurred? (as of Oct 2005)

1. L'Ulivo (Main left-center party)
2. Dialetti della lingua tedesca
3. Magia (Magic)
4. Presidenti a vita
5. Signoraggio
6. Crocifissione di Gesu' (religion)
7. Il presepe napoletano a Maiori
8. Esodo istriano (post-WWII)
9. YHWH (religion)
10. Barbaresco (a top Italian wine)
11. Silvio Berlusconi (our ex-PM)
12. Psicologia
13. Cavalieri templari
14. Internazionale football club
15. Diritto pubblico
16. Foibe (post-WWII massacres)
17. Siena
18. Ischietella
19. Sistema Operativo (CS matters)
20. Hindenburg

Conclusions

- Content-driven reputation:
  - Friendly, does not require user ratings
  - Good prediction (> 80%) of contributions that won't last
- In principle, it works on any versioned document
- Current work:
  - Make it robust to tampering
  - Text trust
  - Social questions

The End

Questions ?