A Content-Driven Reputation System for the Wikipedia

Luca de Alfaro
UC Santa Cruz

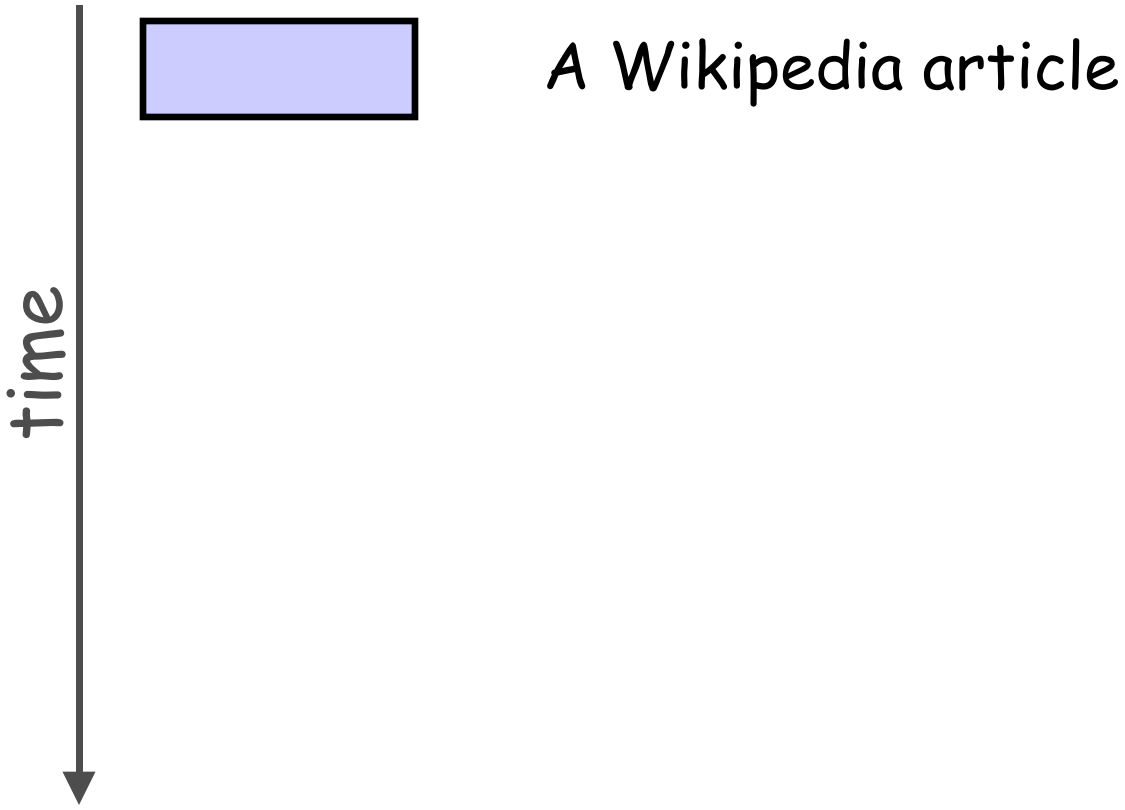
Joint work with Bo Adler, UC Santa Cruz

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.**
  - **Measure author contributions.** It can be used as an incentive for contributing to internal Wikis.

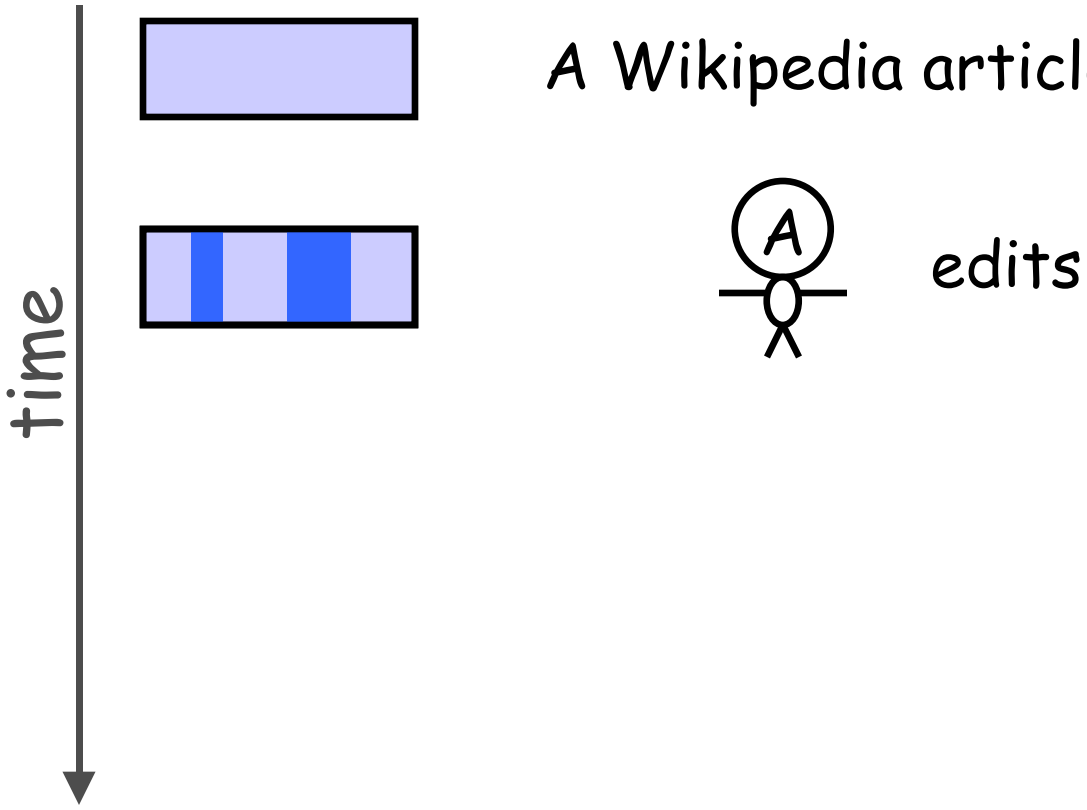
Content-driven reputation

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



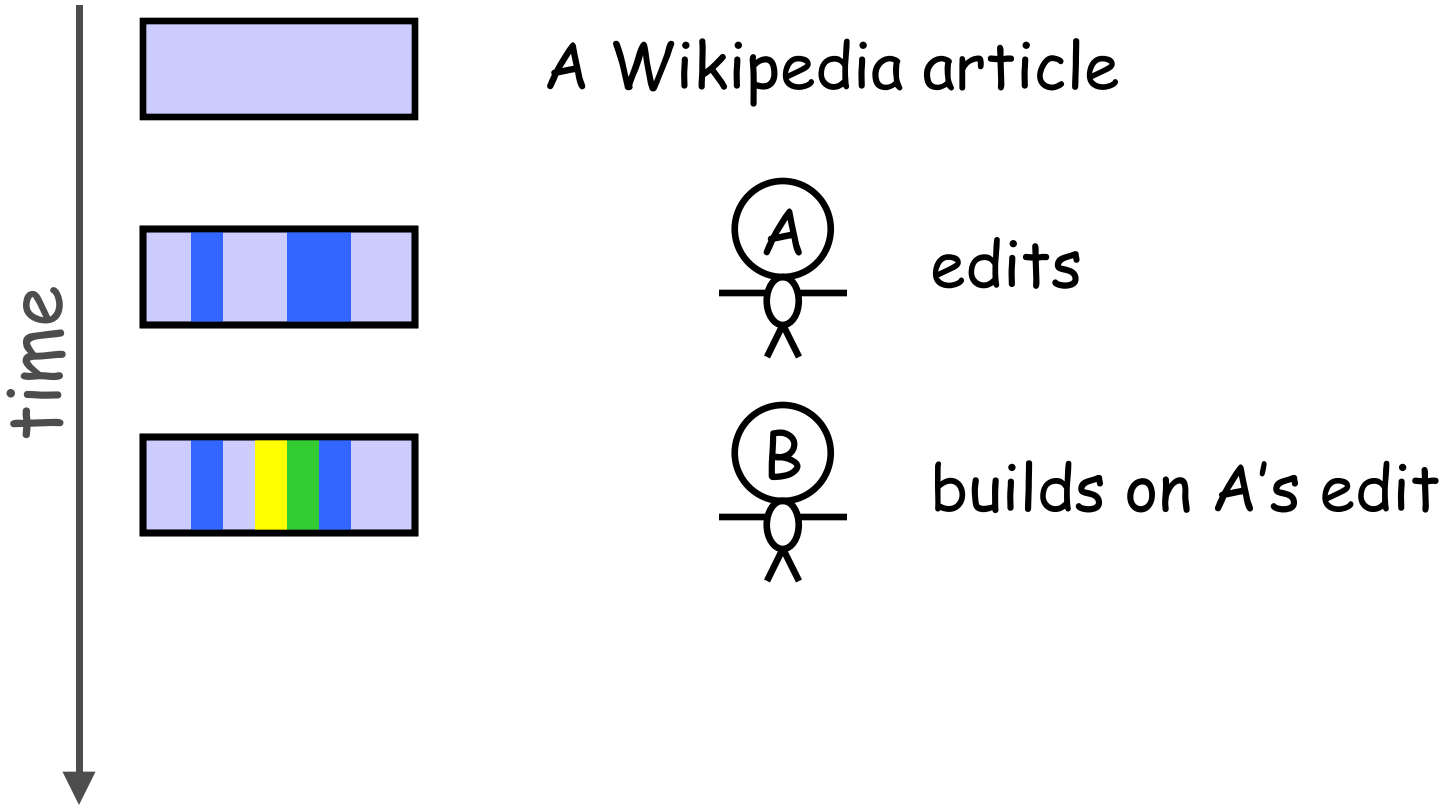
Content-driven reputation

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



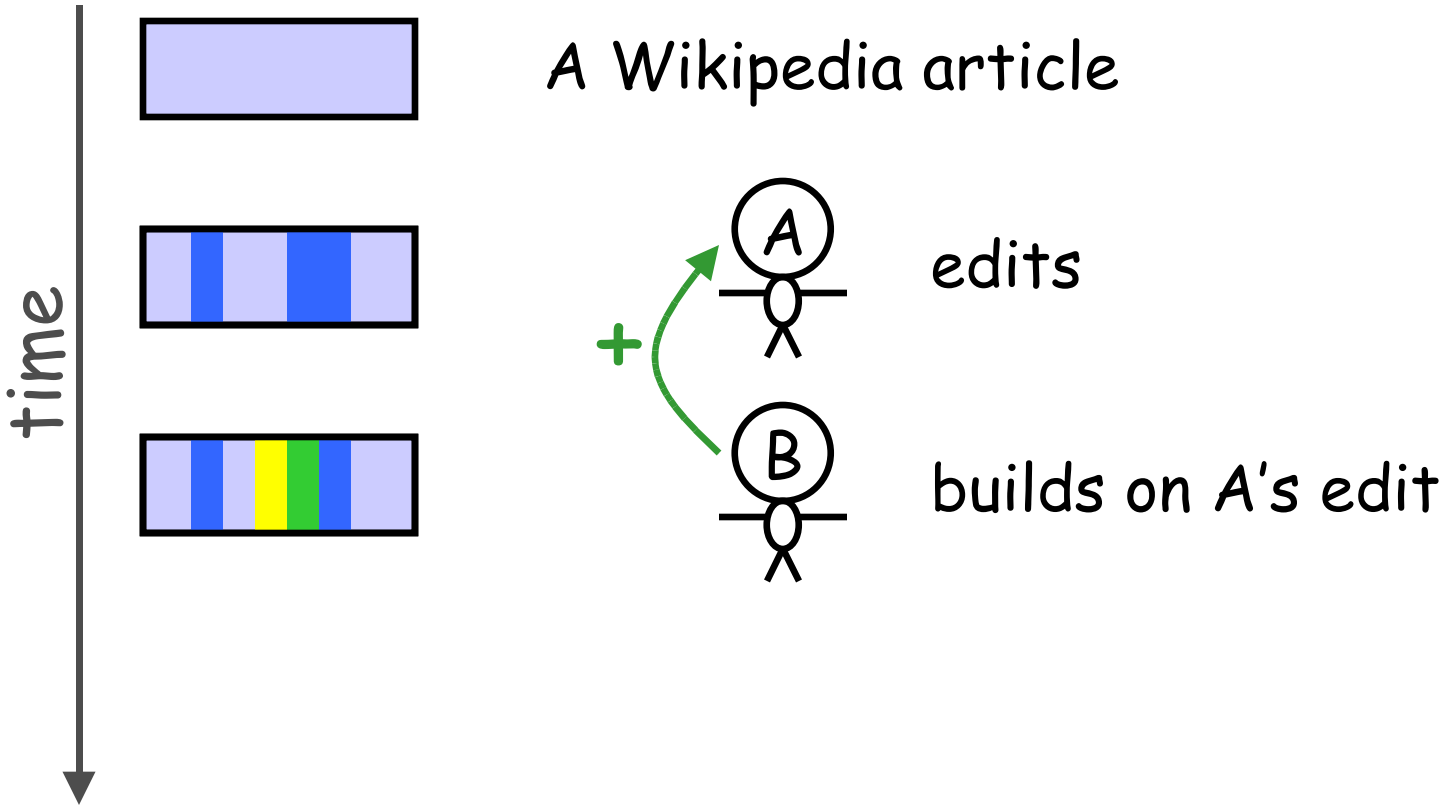
Content-driven reputation

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



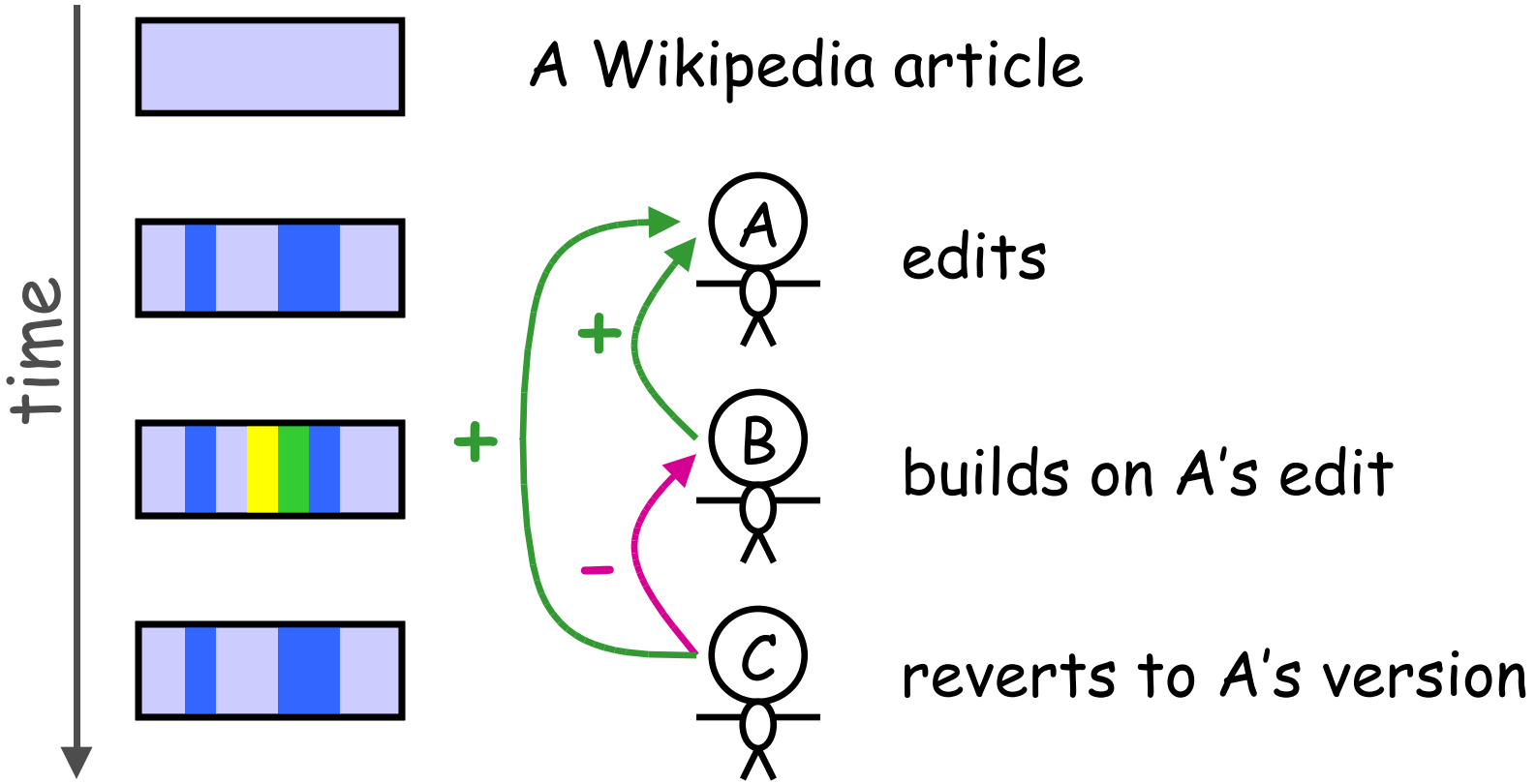
Content-driven reputation

- **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?

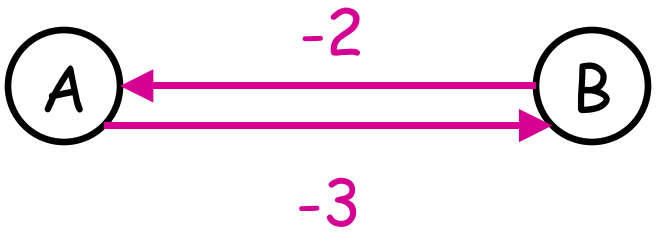
- Many (most?) reputation systems are based on user-to-user comments (e.g., eBay).
- **Content-driven reputation is worth investigating:**
  - **Less social stress: makes people feel more welcome.**
    - No need to explicitly rate others, and be rated.
    - Transparent to casual users (you don't need to know your reputation)
    - 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:

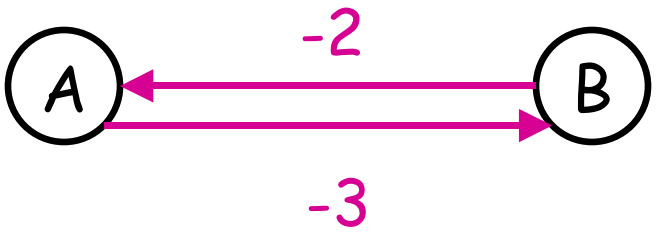
Avoids reputation wars

Wars in user-driven reputation:

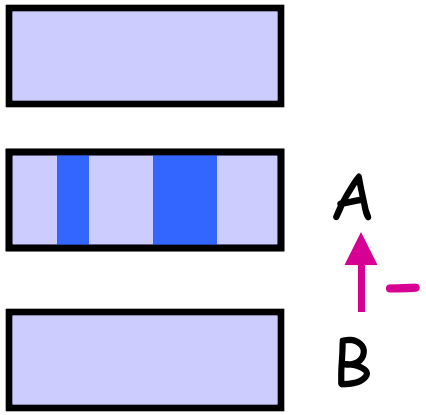


Avoids reputation wars

Wars in user-driven reputation:



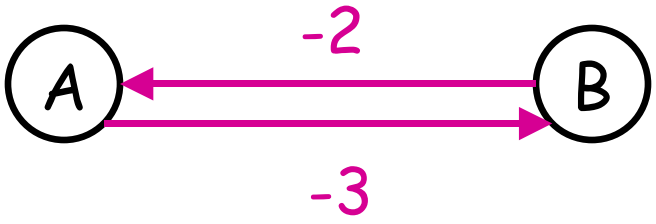
Wars in content-driven reputation:



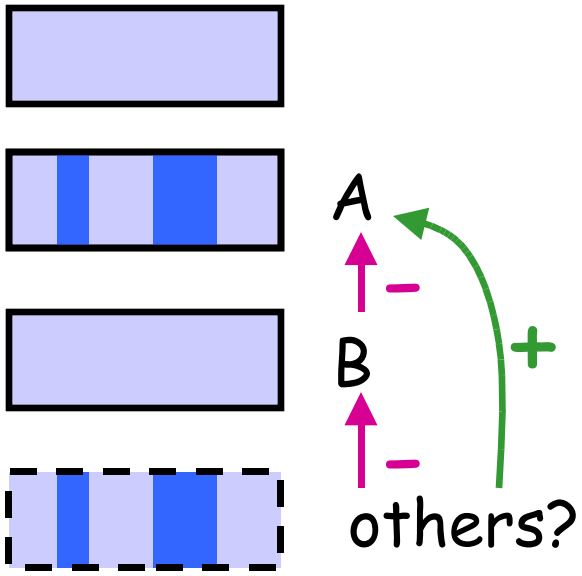
- B can badmouth A by undoing her work

Avoids reputation wars

Wars in user-driven reputation:



Wars in content-driven reputation:

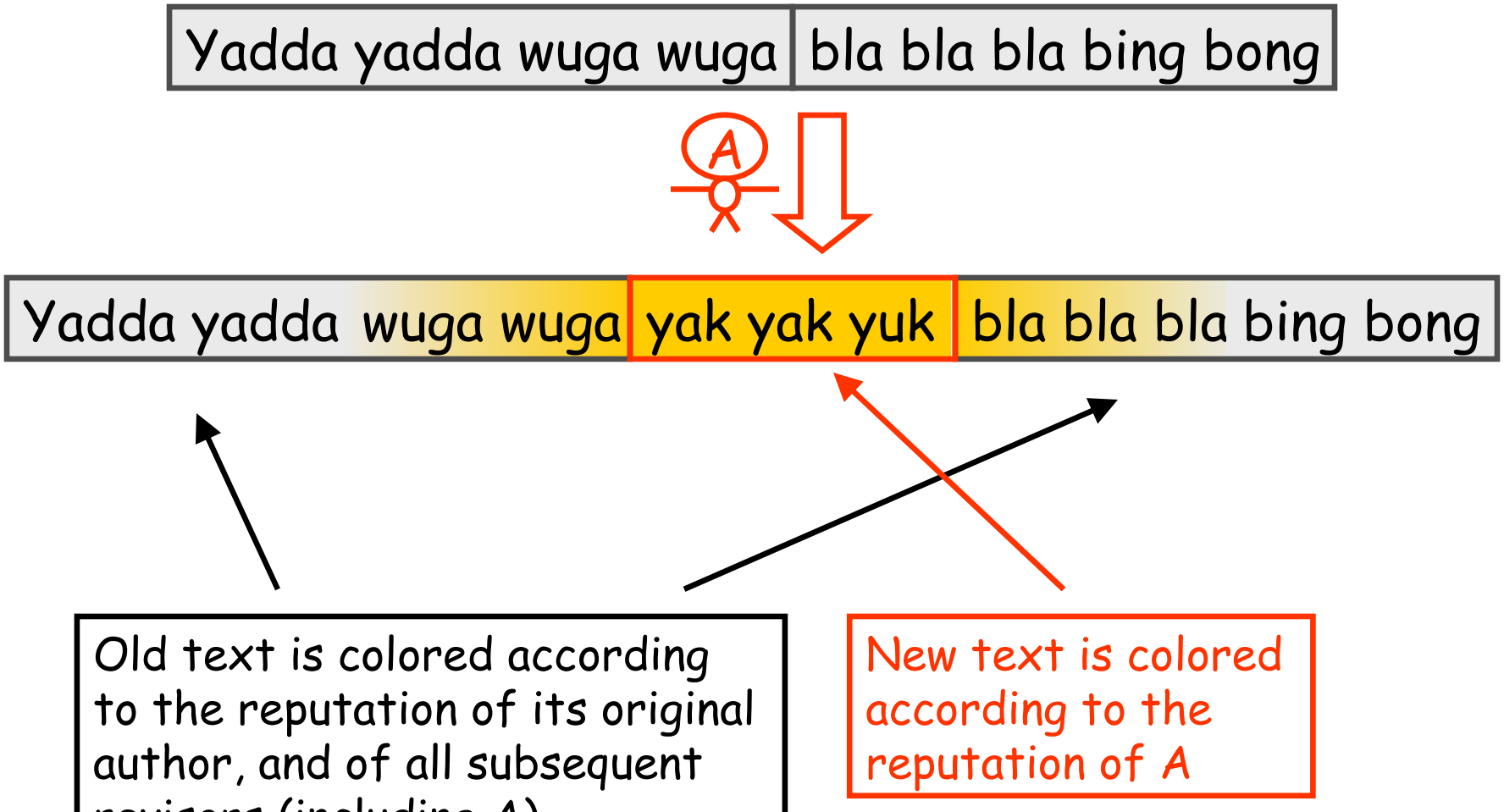


- B can badmouth *A* by undoing her work
- But this is risky: if others then re-instate *A*'s work, it is B's reputation that suffers.

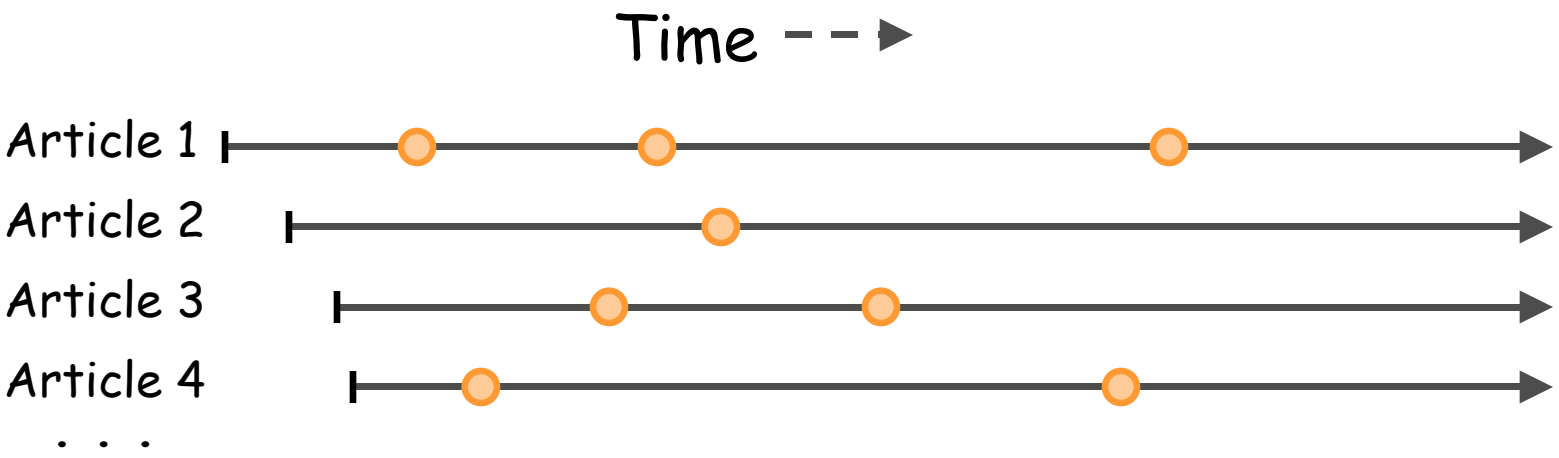
Goals of reputation in Wikipedia

- **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.**

Author reputation and text trust

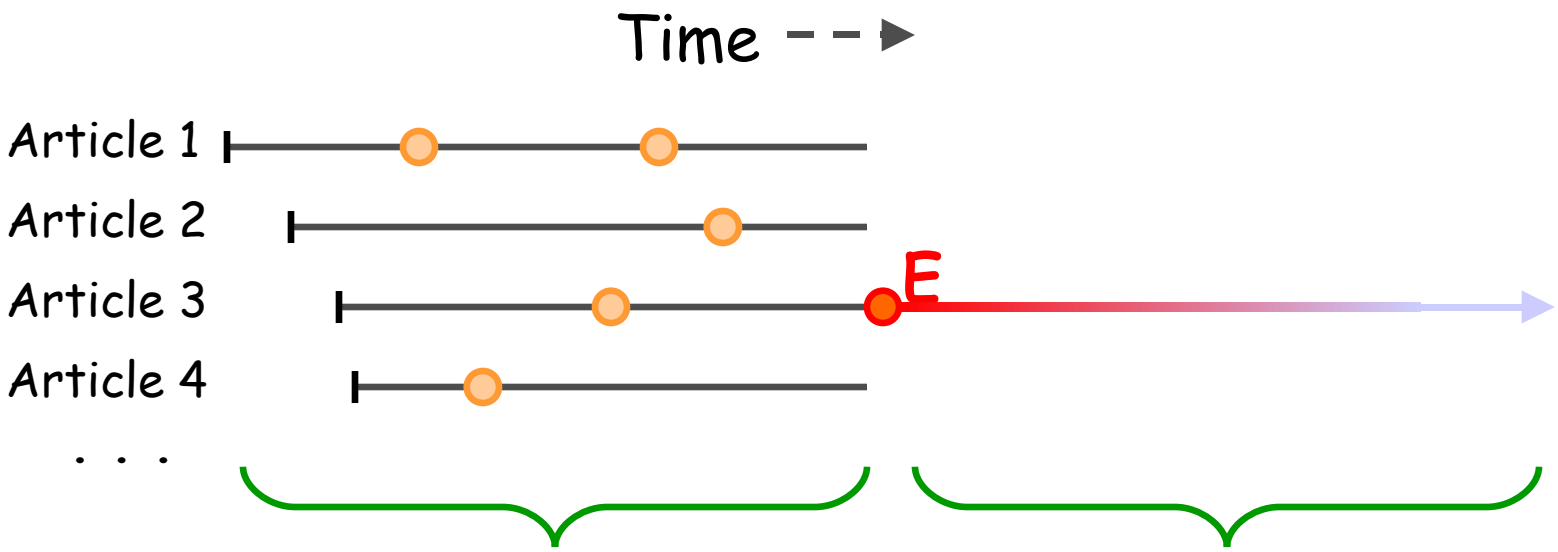


Validation: Does our reputation have predictive value?



= edits by user *A*

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.

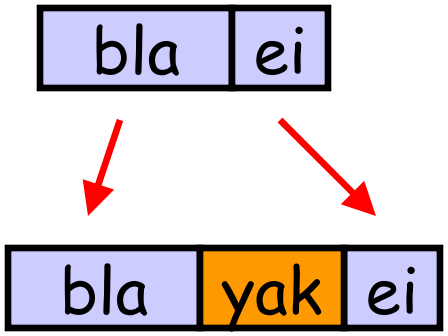
The **longevity** of an edit **E** depends on the history **after** the edit.

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

Building a content-driven reputation system

What is a "contribution"?

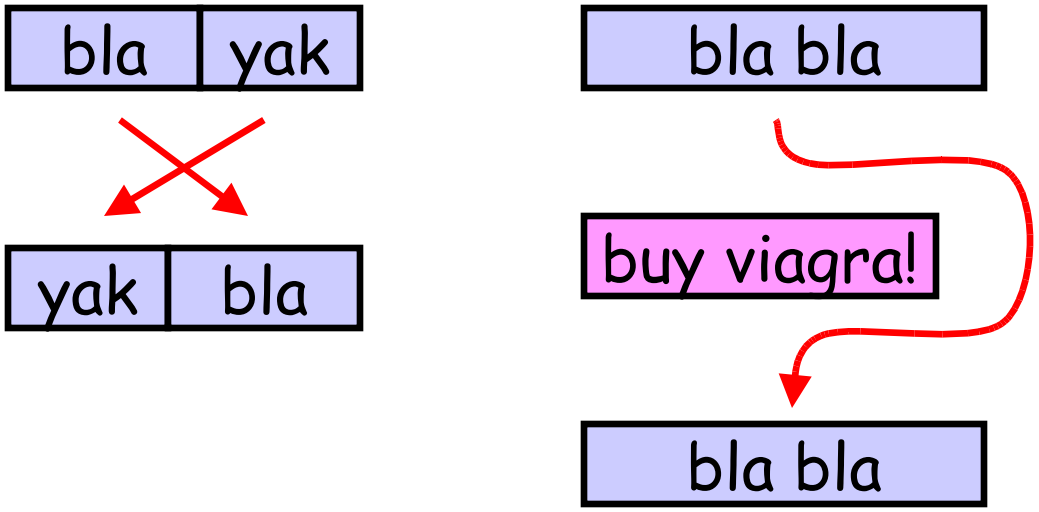
Text



We measure how long the added text survives.

Based on text tracking.

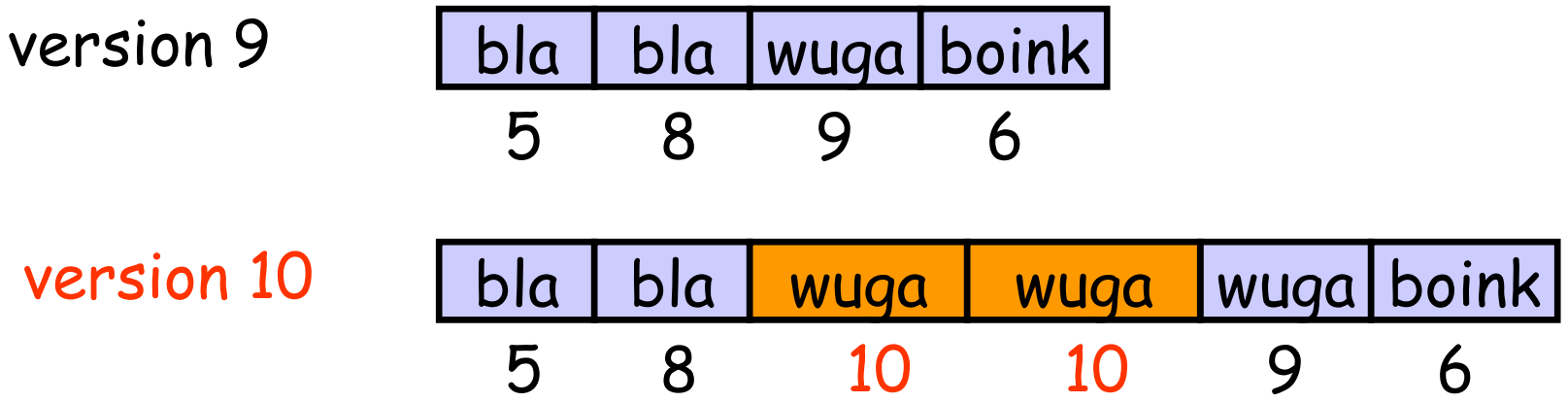
Edit



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

Based on edit distance.

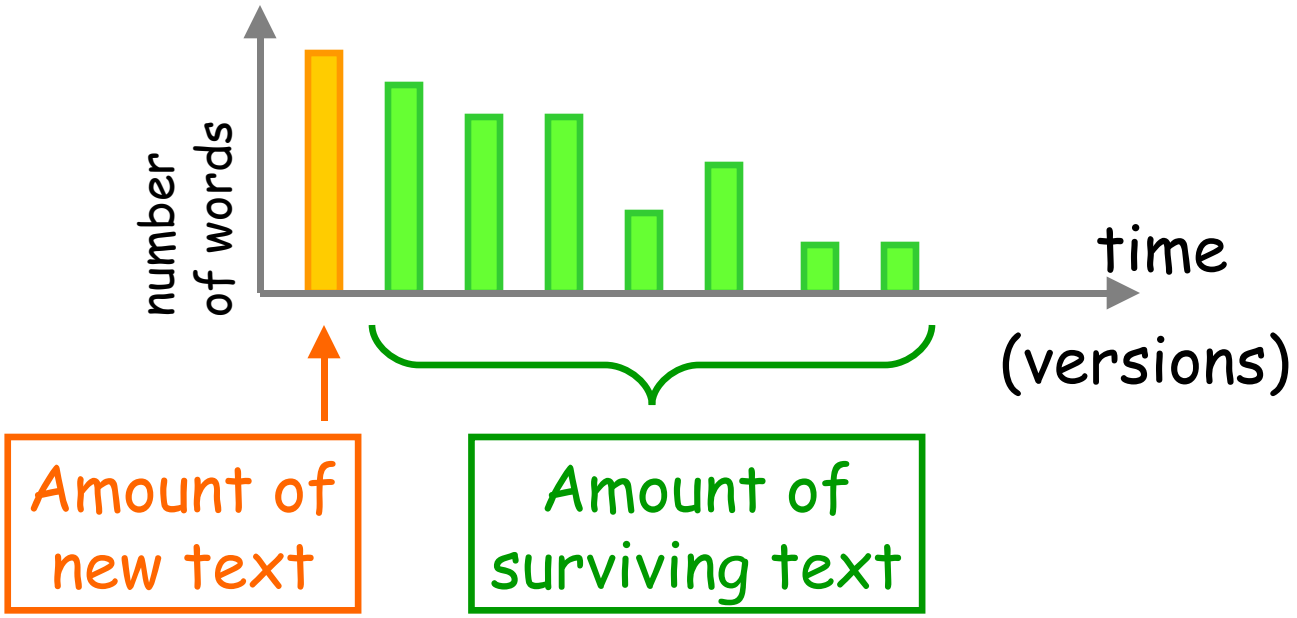
Text



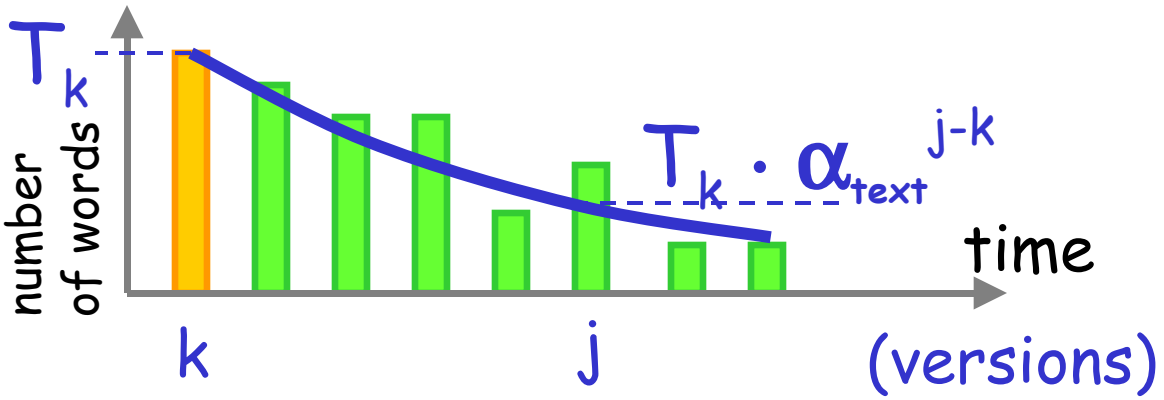
We label each word with the version where it was introduced. This enables us to keep track how long it lives.

(Lots of details omitted; see paper)

Text


number of words
time
(versions)
Amount of new text
Amount of surviving text

Text: Longevity

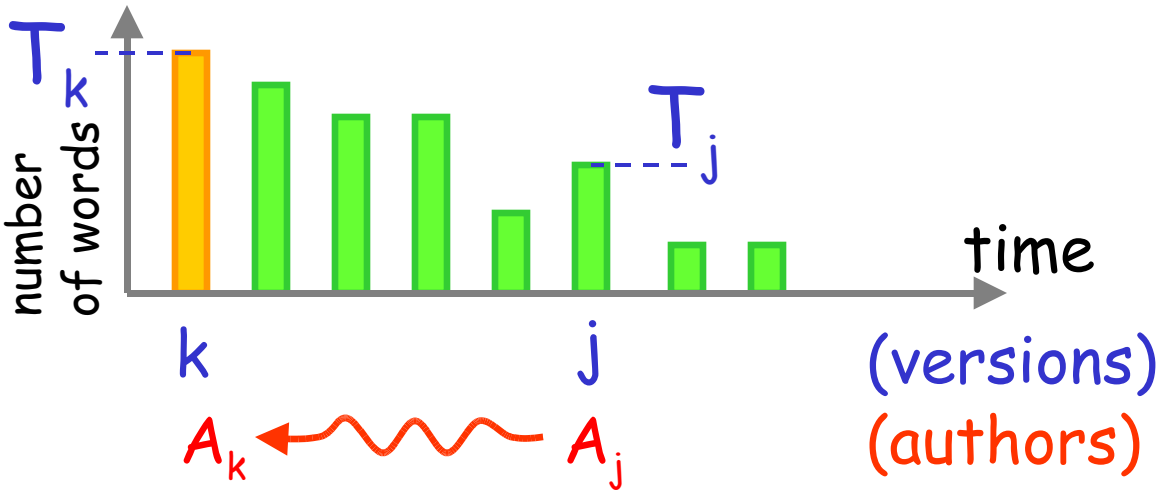


Text Longevity α_{text}:

- 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.

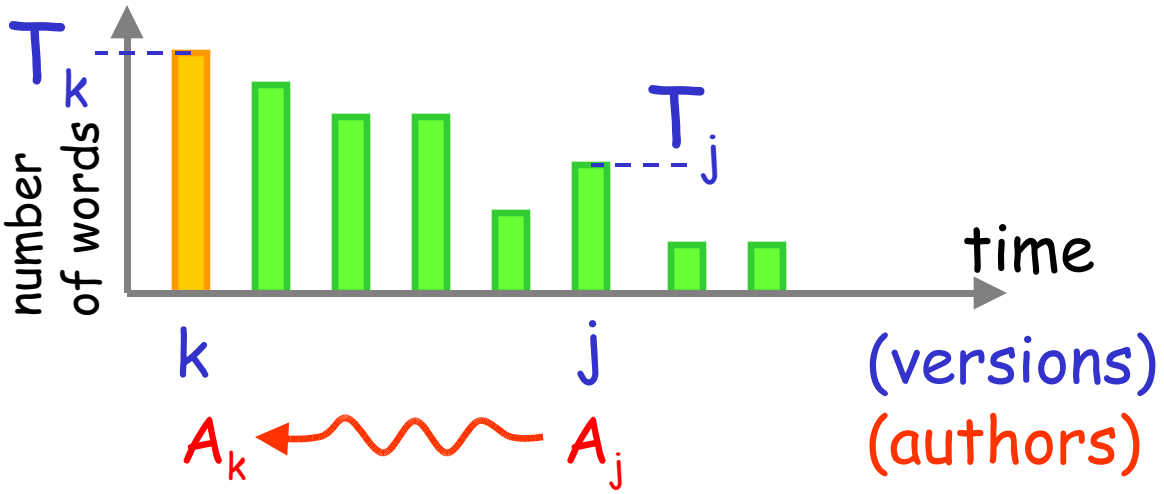
Text: Reputation update



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:

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

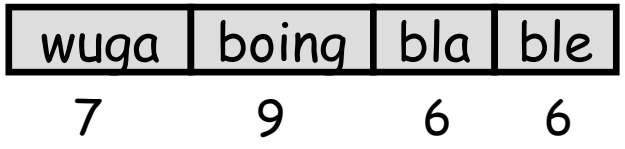
Determined via learning; goal: optimize predictive value.

Measuring text life: Issues

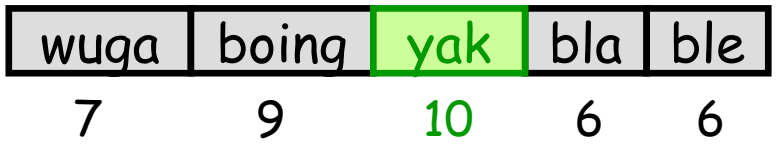
Duplication attack:

Version

9



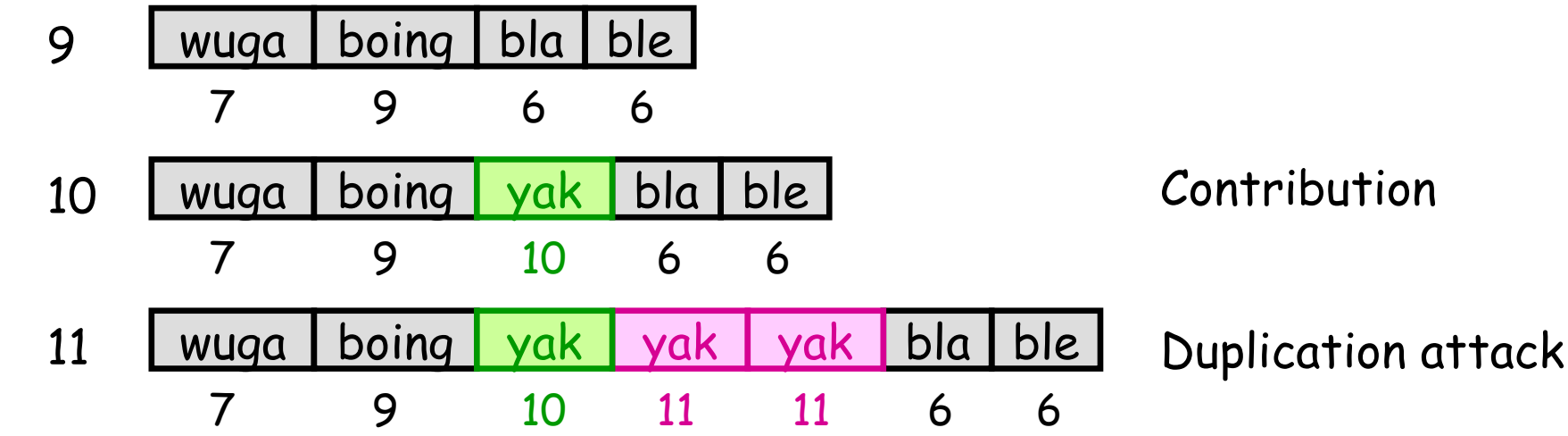
10



Contribution

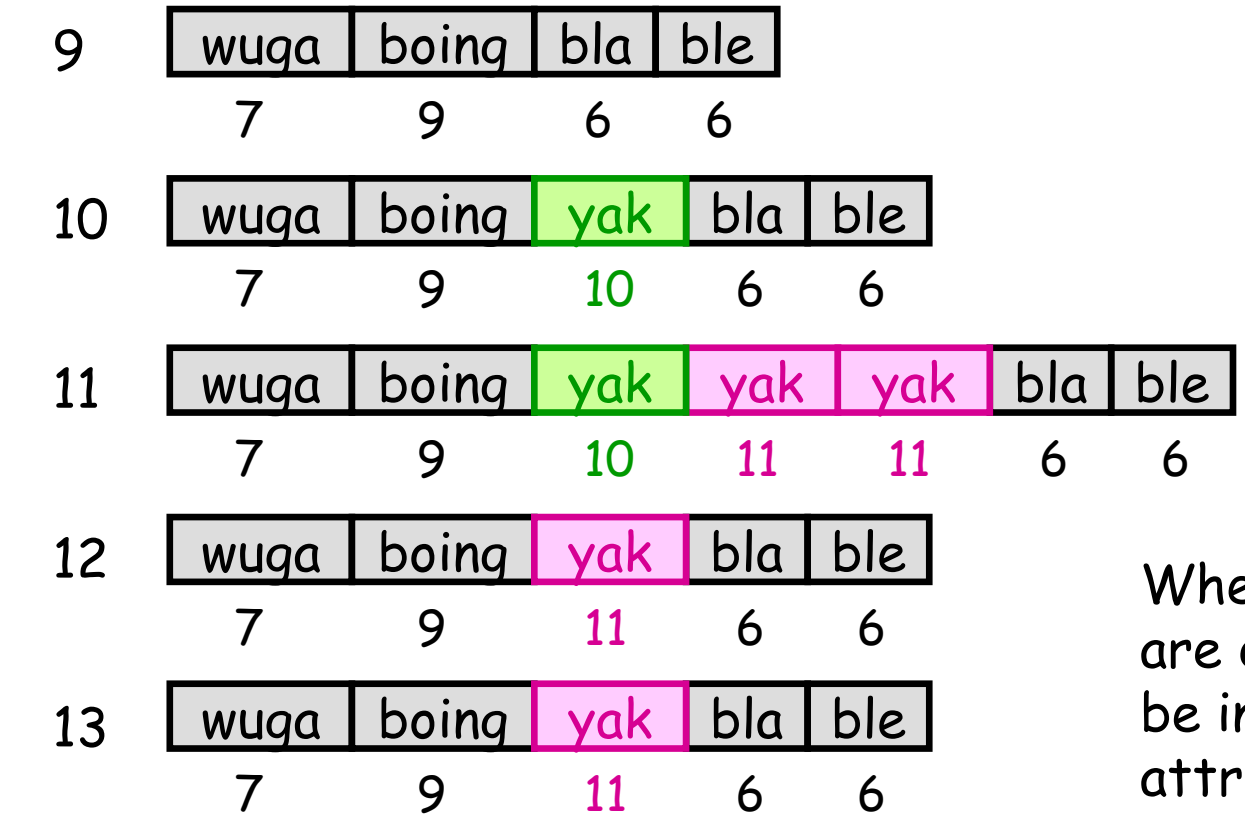
Measuring text life: Issues

Duplication attack:



Measuring text life: Issues

Duplication attack:



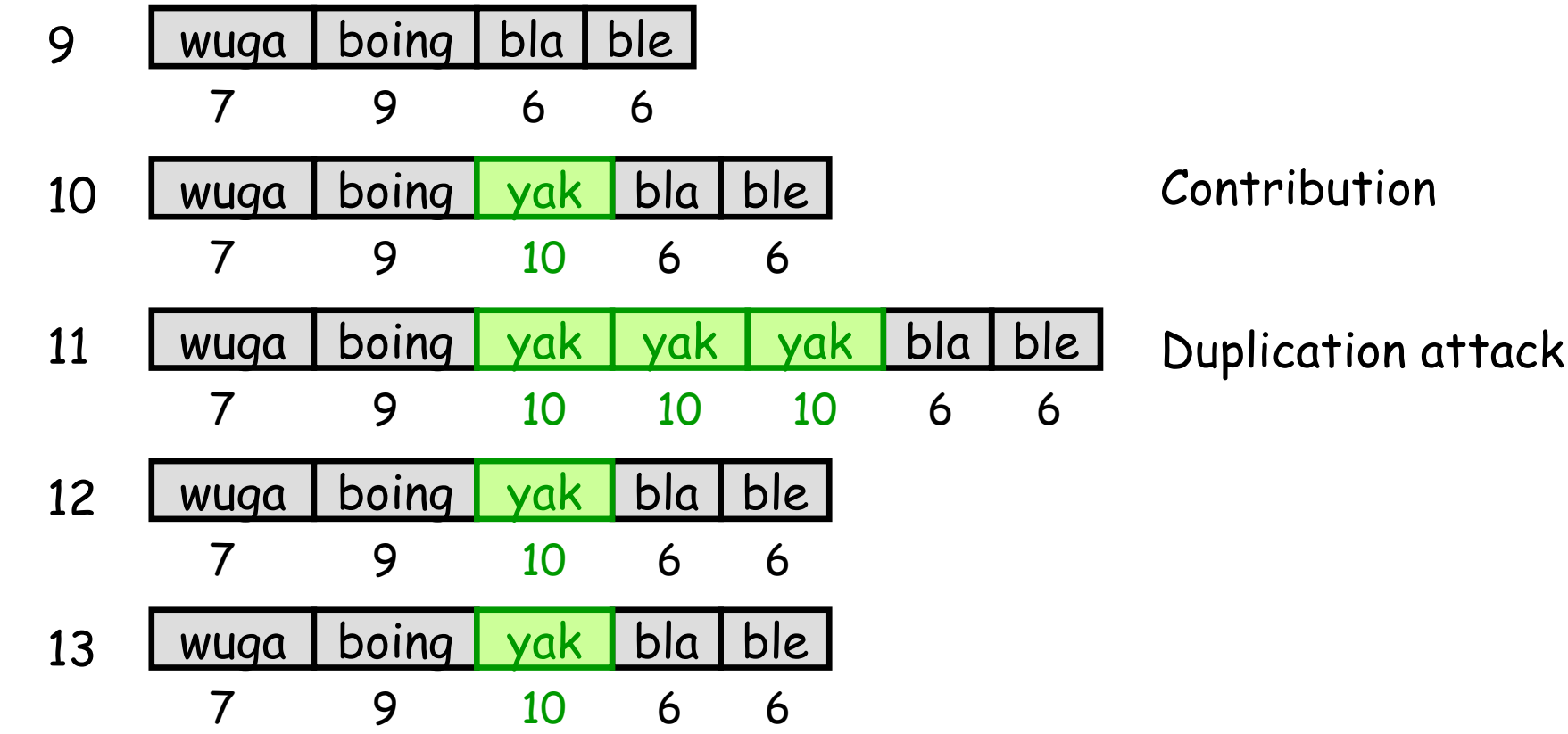
Contribution

Duplication attack

When the duplications are deleted, text may be incorrectly attributed.

Measuring text life: Issues

Duplication attack:

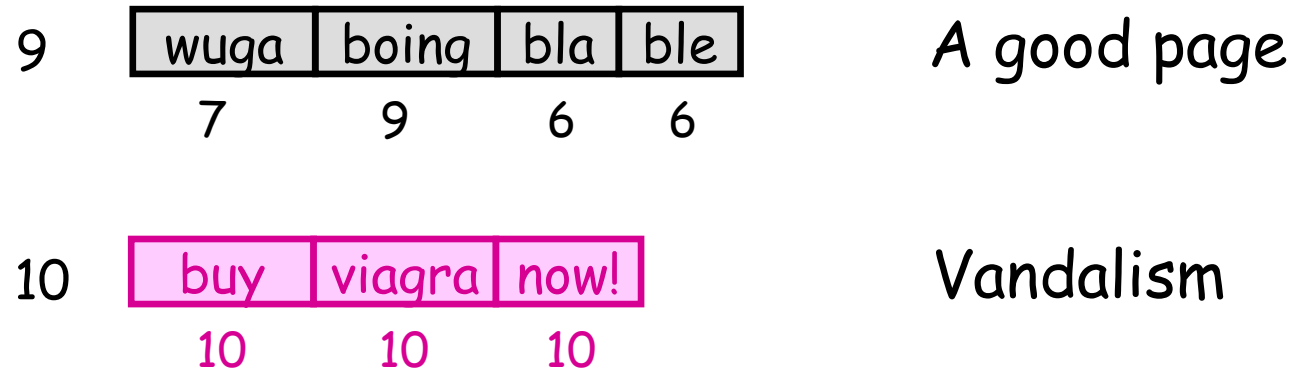


Solution: text can be matched multiple times.

Measuring text life: Issues

Dealing with reversions

Version

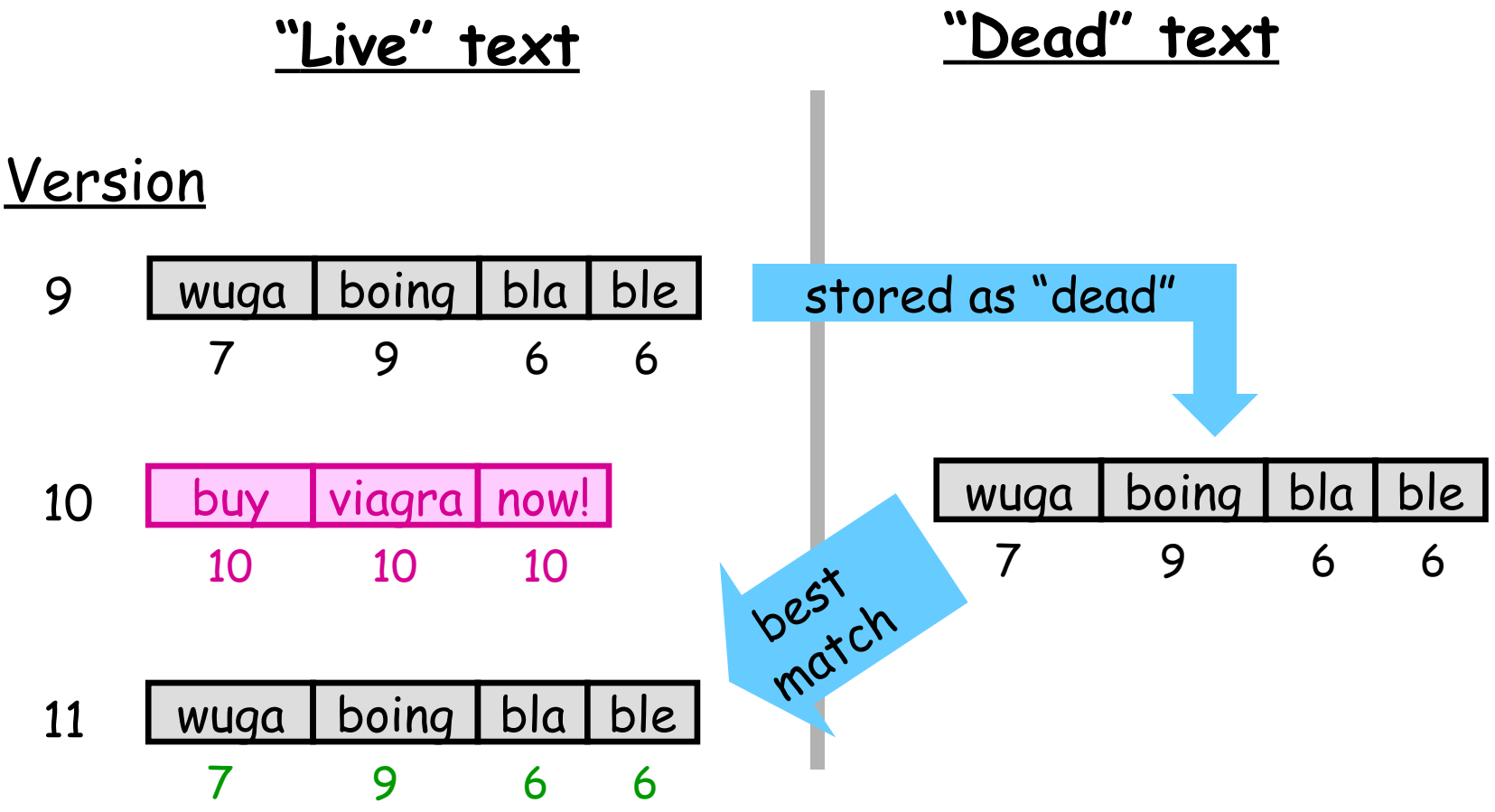


Measuring text life: Issues

Dealing with reversions

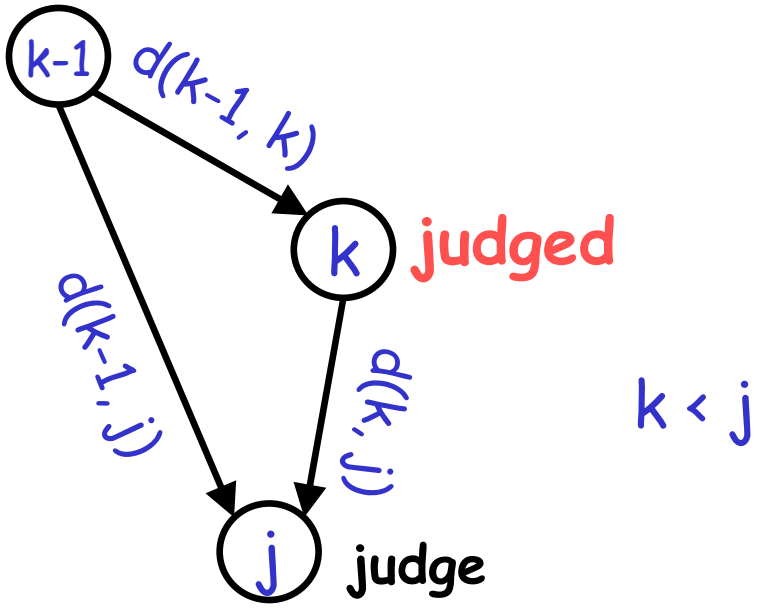
Version

Measuring surviving text



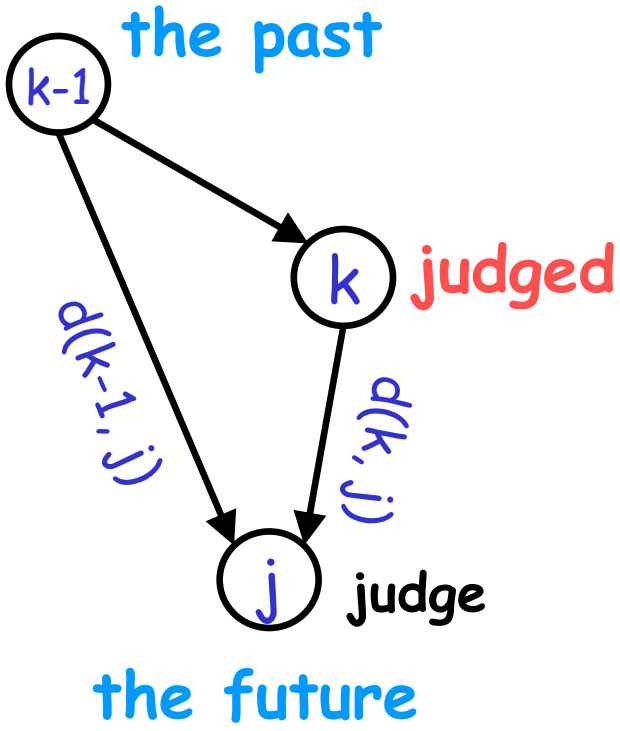
We track authorship of deleted text, and we match the text of new versions both with live and with dead text.

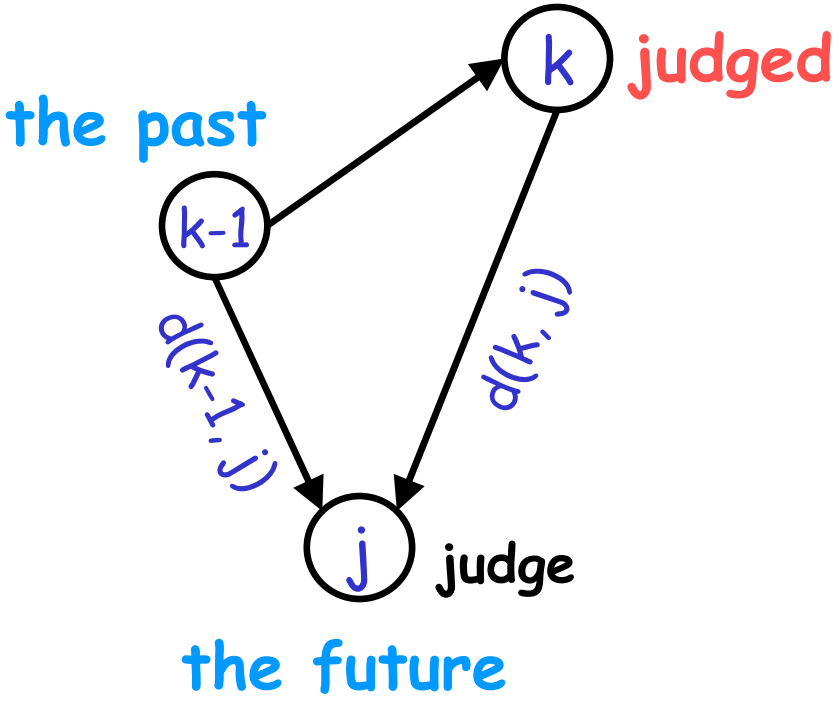
Edit



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?





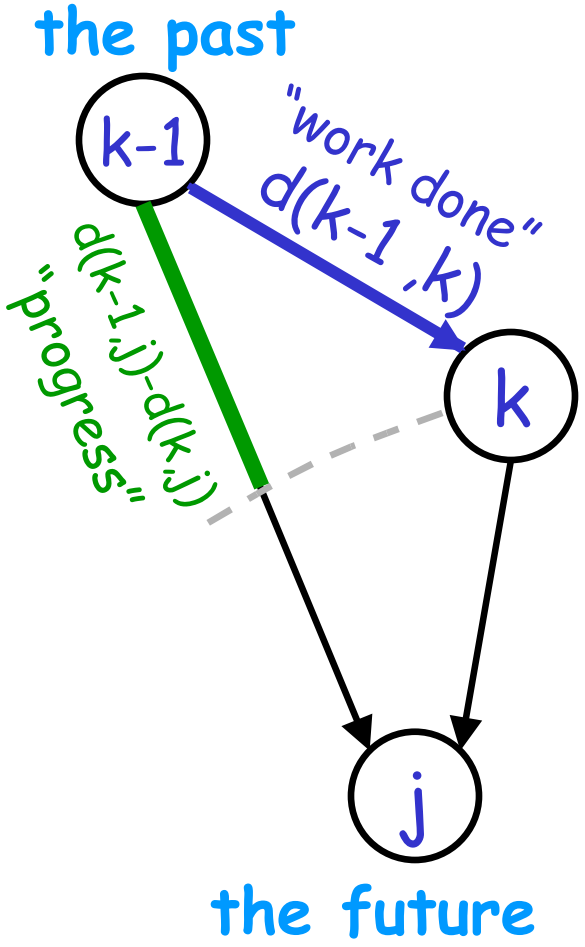
k is **good**: d(k-1, j) > d(k, j)

"k went towards the future"

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

"k went against the future"

Edit: Longevity



Edit Longevity:

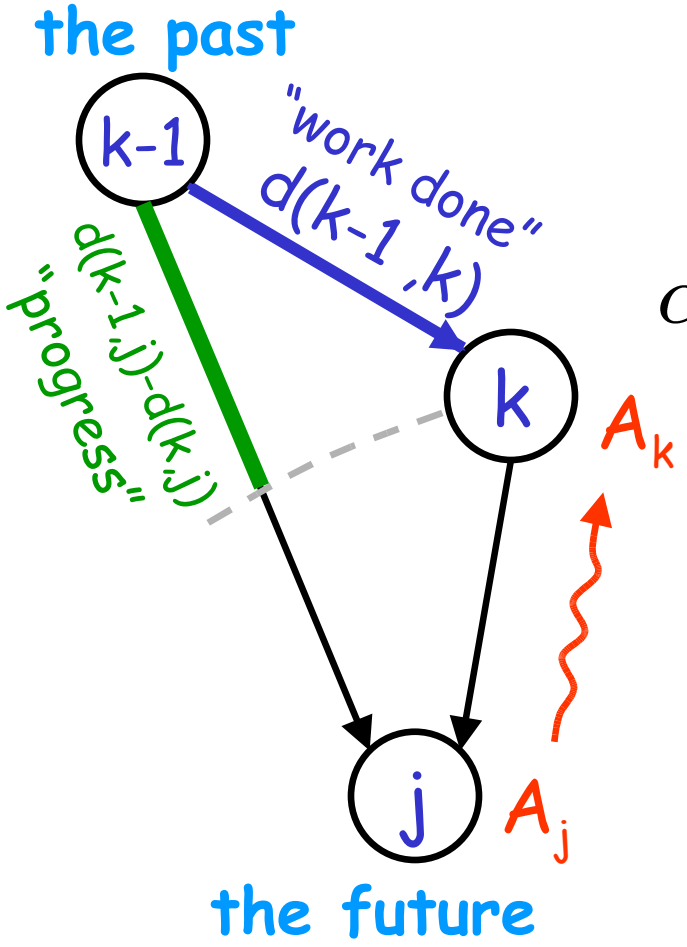
$$\alpha_{edit} = \frac{d(k-1,j) - d(k,j)}{d(k-1,k)}$$

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:

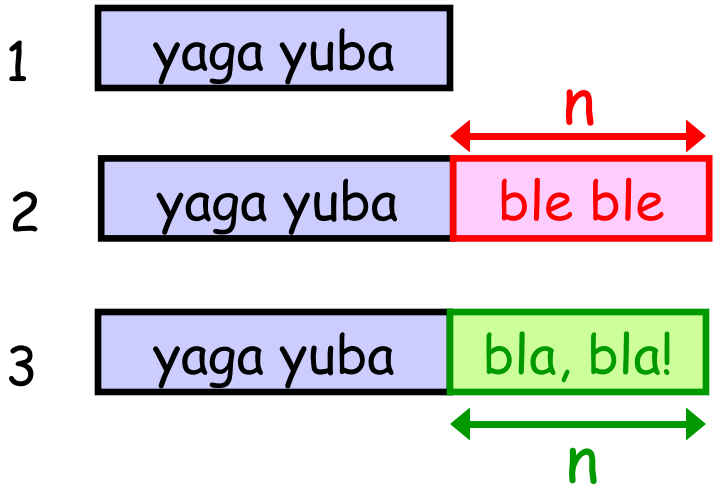
- I_{tot}: Total inserted text
- D_{tot}: Total deleted text
- M_{tot}: Total amount of relative exchange of position (a measure of text reordering).

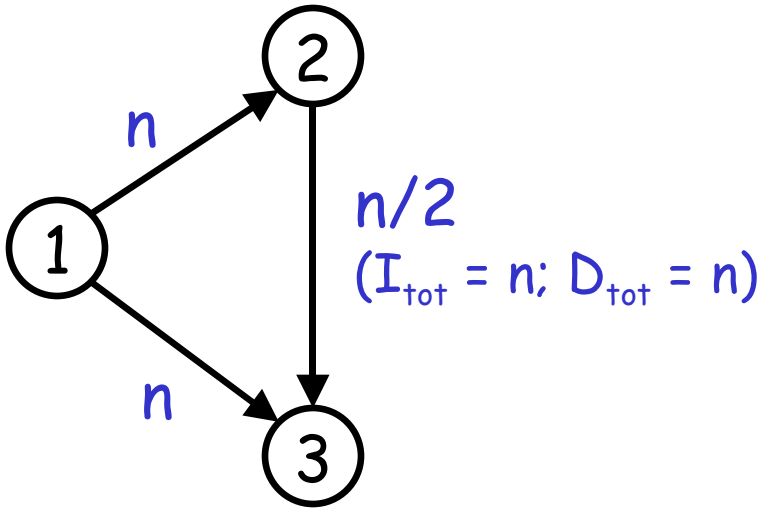
$$d(u,v) = \max(I_{tot}, D_{tot}) - \frac{1}{2}\min(I_{tot}, D_{tot}) + M_{tot}$$

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 extensive 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 extensive 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

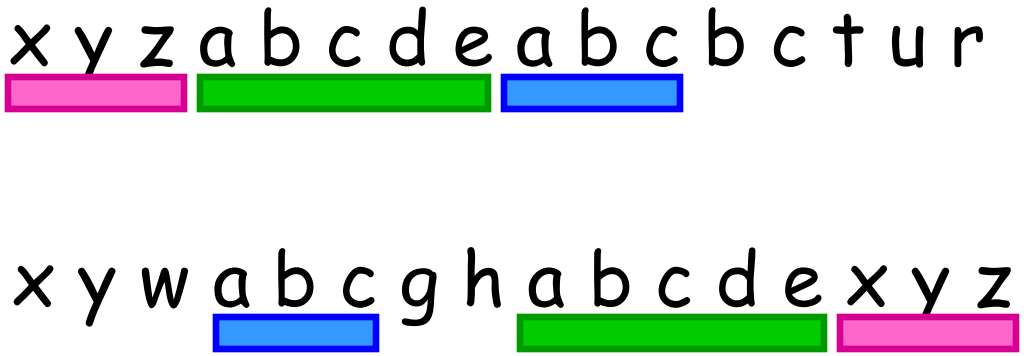
x y z a b c d e a b c b c t u r

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

After extensive 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 extensive 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 extensive 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 power

- We learned the coefficients in our reputation update formulas, with the goal of maximizing the predictive value of reputation (recall x precision, mutual information, ...)
- Training set: Italian Wikipedia (till Oct 05)
  154,261 pages; 714,280 versions
- Evaluation set: French Wikipedia (till Jul 06)
  536,930 pages; 4,873,243 versions
- With an improved code-base, we can now handle the English Wikipedia with a single PC.

Results: French Wikipedia, in detail

% of edits below a given longevity

log (1 + reputation)

Bin	%_data	1.0	0.6	0.2	-0.2	-0.6	-1.0
0	34.59	100	56.23	53.72	52.41	50.77	25.80
1	0.6	100	17.44	14.40	13.08	11.43	4.13
2	1.07	100	17.71	14.41	12.94	11.38	3.26
3	1.73	100	13.23	10.27	8.84	7.23	2.73
4	2.53	100	13.25	9.94	8.28	6.73	2.06
5	4.23	100	12.40	9.88	8.62	6.81	2.25
6	5.3	100	12.01	9.01	7.52	5.99	2.21
7	8.04	100	13.67	9.77	7.99	6.15	1.59
8	9.91	100	12.51	9.65	8.10	6.33	0.79
9	32	100	11.38	8.50	7.07	5.39	1.62

Results: French Wikipedia, in detail

% of edits below a given longevity

Bin	% data	1.0	0.6	0.2	-0.2	-0.6	-1.0
0	34.59	100	56.23	53.72	52.41	50.77	25.80
1	0.6	100	17.44	14.40	13.08	11.43	4.13
2	1.07	100	17.71	14.41	12.94	11.38	3.26
3	1.73	100	13.23	10.27	8.84	7.23	2.73
4	2.53	100	13.25	9.94	8.28	6.73	2.06
5	4.23	100	12.40	9.88	8.62	6.81	2.25
6	5.3	100	12.01	9.01	7.52	5.99	2.21
7	8.04	100	13.67	9.77	7.99	6.15	1.59
8	9.91	100	12.51	9.65	8.10	6.33	0.79
9	32	100	11.38	8.50	7.07	5.39	1.62

low rep

log (1 + reputation)

Short-Lived

Results: defining the terms

Based on our system:

- **Low-reputation:** Lower 20% of range
- **Reversions:** $\alpha_{edit} \leq -0.8$ (almost entirely undone)
- **Short-lived text:** $\alpha_{text} \leq 0.2$ (disappers quickly)

Based on a user study:

We asked 6 people to rate with {-1, 0, +1} 680 revisions of the Italian Wikipedia. Each revision was rated three times. The data is noisy (people often disagree!).

- **Bad text, bad edit:** score ≤ -0.33 (lower third of range)

Predictive power of low reputation

Excluding Anonymous Authors

Predicting reversions

Precision	23.9 %
Recall	32.2 %

Predicting short-lived text

Precision	5.9%
Recall	37.8%

Predicting bad edits

Recall	49 %

*

Predicting bad text

Recall	79 %

*

Edits by low-reputation users are 4.2 times more likely than average to be reverted.

* On Italian Wikipedia

Predictive power of low reputation

Excluding Anonymous Authors

Predicting reversions

Precision	23.9 %
Recall	32.2 %

Predicting short-lived text

Precision	5.9%
Recall	37.8%

Including Anonymous Authors

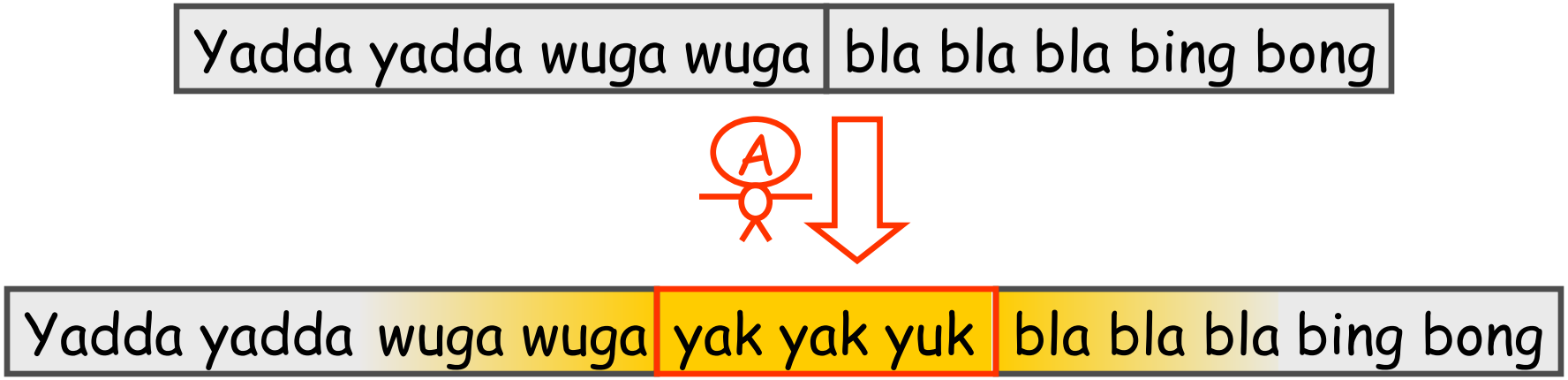
Predicting reversions

Precision	48.9 %
Recall	82.9 %

Predicting short-lived text

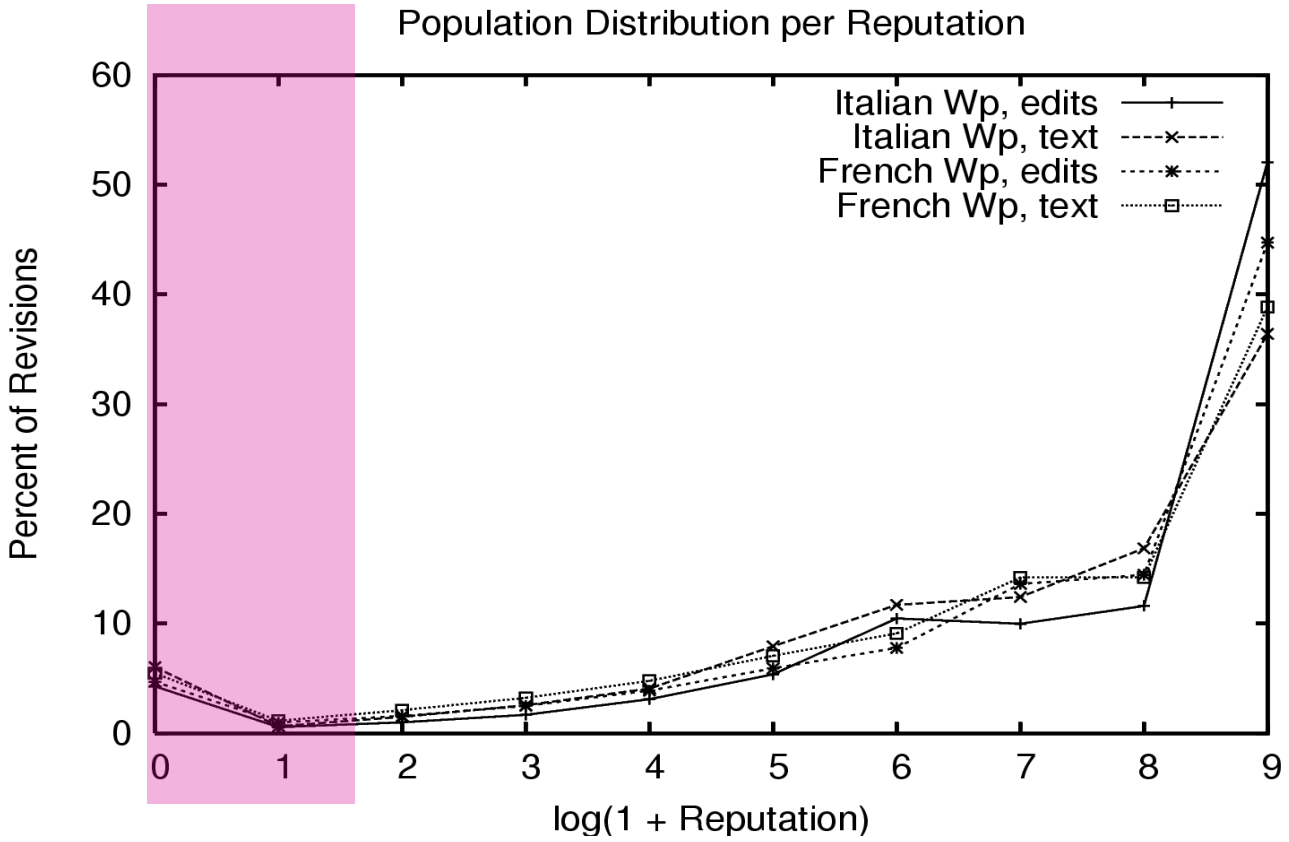
Precision	19.0 %
Recall	90.4 %

Author reputation and text trust



- If we color text fresh from low-reputation authors, we color over 80% of text which will be short-lived!
- We can spot which text won't make it!
- Text trust is still work in progress...

Distribution of authors by reputation



Low reputation

The data excludes anonymous users.

We can extract a fairly complete picture of social interaction on the Wikipedia

- We can automatically detect reversions, edit wars, and other phenomena!
- A student is looking at the following question: is topic-specific reputation a better predictor of edit quality than global reputation?
- We can measure these things!! No need to "guess".
- This opens the way to all sort of sociological studies of the Wikipedia, and of how people collaborate.

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 (medioeval topic)
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:
  - Can be given automatically
  - Friendly, does not require user ratings
  - Good recall (> 80%) of contributions that won't last
  - Similar to edit count if no gaming takes place
- In principle, it works on any versioned document
- Current work:
  - Make it robust to tampering
  - Text trust

Questions?