Streaming Algorithms for Strings with Mismatches

Funda Ergün[1], Elena Grigorescu[2], Erfan Sadeqi Azer[1], Samson Zhou[2]

[1]Indiana University, [2]Purdue University

samsonzhou@gmail.com

PRELIMINARIES

❖ Palindrome: A string S that reads the same forwards and backwards, $S = S^R$ (Ex: RACECAR)

❖ Period: The length of a substring which is continuously repeated in a string S (Ex: abcabcabcabc)

❖ What if there are errors in the data?

❖ d-near-palindromes: Given a metric $dist$, a d-near-palindrome has $dist(S, S^R) \leq d$.

- Use Hamming distance for metric
- Ex: FACECAR for $d = 2$.

❖ k-period: A string S has k-period p if and only if $\text{HAM}(S[1, n-p], S[p+1, n]) \leq k$.

- (Ex: abcaacaadaad) for k = 2.

Questions:
Given a data stream S, can we find the smallest k-period of S and the longest d-near-palindrome contained in S ?

Properties of Karp-Rabin Fingerprints:

❖ Given base B and a prime P, define $\phi(S) = \sum_{i=1}^{n} B^i S[i] \ (mod\ P)$

❖ If $S = T$, then $\phi(S) = \phi(T)$

❖ If $S \neq T$, then $\phi(S) \neq \phi(T)$ w.h.p. (Schwartz-Zippel)

❖ $\phi(S[1{:}y]) = \phi(S[1{:}x]) + B^x \phi(S[x{:}y])$ (sliding)

❖ Define $\phi^R(S) = \sum_{i=1}^{n} B^{-i} S[i] \ (mod\ P)$

❖ $\phi(S^R[1{:}x]) = B^{x+1} \phi^R(S[1{:}x])$ (reversal)

❖ $\phi^R(S[1{:}y]) = \phi^R(S[1{:}x]) + B^{-x} \phi^R(S[x{:}y])$

RELATED WORK

❖ $O(\log n)$ space to provide a $(1+\varepsilon)$ multiplicative approximation to the length of the longest palindrome (BEMS14)

❖ $O(\sqrt{n})$ space to provide a $\sqrt{n}$ additive approximation to the length of the longest palindrome (BEMS14)

❖ $O(\sqrt{n})$ space to find the longest palindrome in two passes (BEMS14)

❖ $\Omega\left(\frac{\log n}{\varepsilon \log(1+\varepsilon)}\right)$ space for $(1+\varepsilon)$ multiplicative approximation (GMSU16)

❖ $\Omega\left(\frac{n}{E}\right)$ space for E additive approximation (GMSU16)

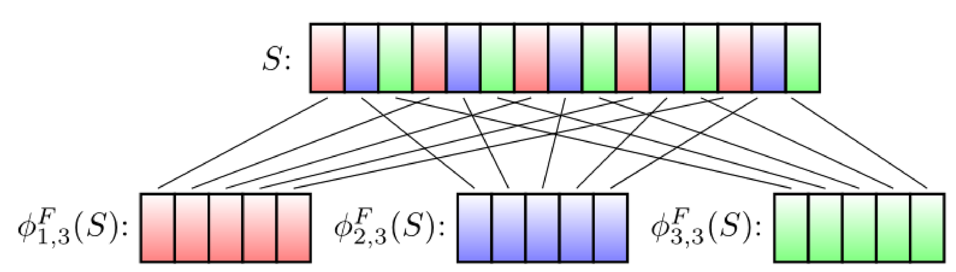
❖ $O(\log^2 n)$ space to find the shortest period in one-pass (EJS10)

❖ $\Omega(n)$ space to find the period, if aperiodic, in one-pass. (EJS10)

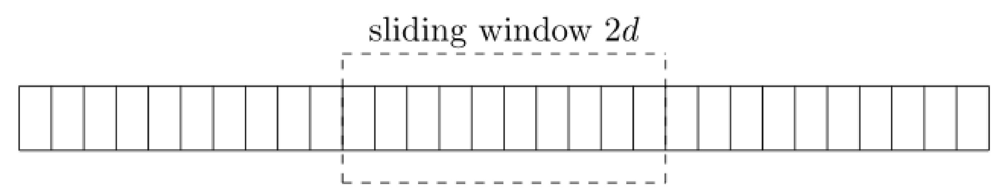
❖ $O(\log^2 n)$ space to find the shortest period in two-passes, even if aperiodic (EJS10)

TECHNIQUES (NEAR-PALINDROMES)

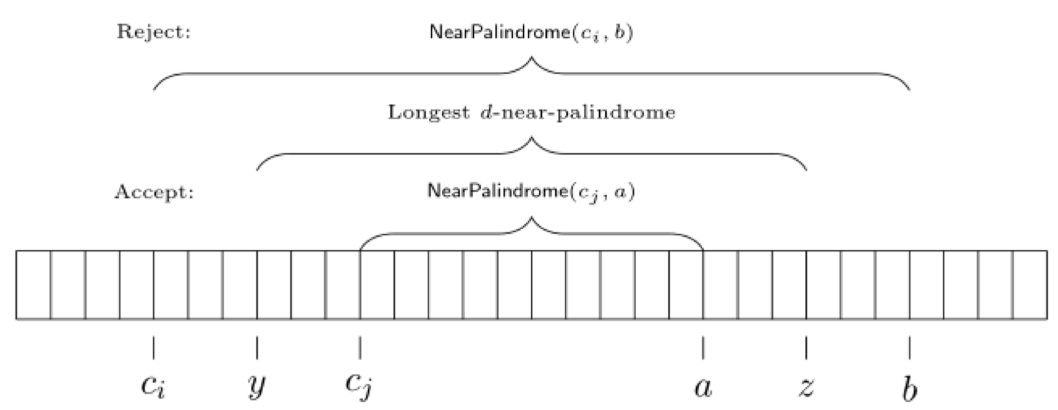
$$S_{a,b} = S[a]S[a+b]S[a+2b]S[a+3b]\ldots$$
$$\phi_{a,b}(S) = \phi(S_{a,b})$$
$$= B * S[a] + B^2 * S[a+b] + B^3 * S[a+2b]\ldots$$



Maintain a sliding window of size $2d$ to find all short d-near-palindromes.



Dynamically maintain a series of checkpoints, and see if the substrings are d-near-palindromes.



RESULTS (NEAR-PALINDROMES)

❖ $O\left(\frac{d \log^7 n}{\varepsilon \log(1+\varepsilon)}\right)$ space to provide a $(1+\varepsilon)$ multiplicative approximation to the length of the longest d-near-palindrome

❖ $O(d\sqrt{n}\log^6 n)$ space to provide a $\sqrt{n}$ additive approximation to the length of the longest d-near-palindrome

❖ $O(d^2\sqrt{n}\log^6 n)$ space to find the longest d-near-palindrome in two passes

❖ $\Omega(d \log n)$ space LB for $(1+\varepsilon)$ multiplicative approximation

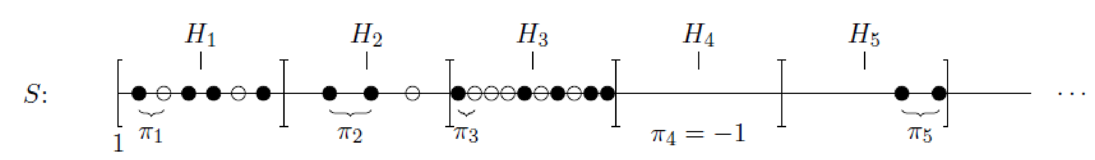
❖ $\Omega\left(\frac{dn}{E}\right)$ space LB for E additive approximation

	Longest Palindrome	Longest d-Near-Palindrome
$(1+\varepsilon)$ multiplicative	$O(\log^2 n)$ (BEMS14)	$O\left(\frac{d \log^7 n}{\varepsilon \log(1+\varepsilon)}\right)$
$\sqrt{n}$ additive	$O(\sqrt{n}\log n)$ (BEMS14)	$O(d\sqrt{n}\log^6 n)$
two pass exact	$O(\sqrt{n}\log n)$ (BEMS14)	$O(d^2\sqrt{n}\log^6 n)$
$(1+\varepsilon)$ multiplicative LB	$\Omega\left(\frac{\log n}{\log(1+\varepsilon)}\right)$ (GMSU16)	$\Omega(d \log n)$
E additive LB	$\Omega\left(\frac{n}{E}\right)$ (GMSU16)	$\Omega\left(\frac{dn}{E}\right)$

TECHNIQUES (K-PERIODICITY)

First pass: Find all indices i at which a substring which is a near-match to $S\left[1, \frac{n}{2}\right]$ begins.

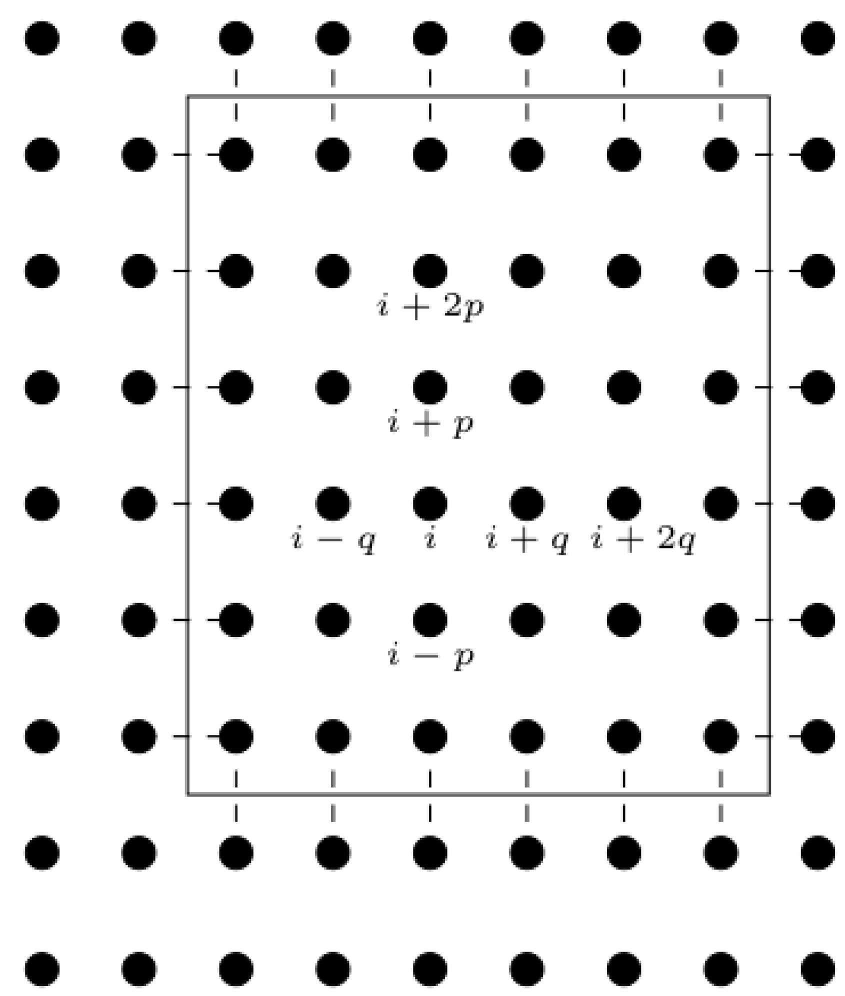
First pass: Add in a small number of false positives to allow easy compression



Observe that all dots in each interval are equally spaced after the first. The black dots are the list of candidates, while the white dots are false positives that we include to allow easy compression.

What can we say about two candidates p and q?

If p and q are "small", then $\gcd(p, q)$ is a $O(k^2)$-period.



The dashed lines are bad edges. The total area of the enclosed regions can be at most k^2 if the perimeter is at most $4k$.

LOWER BOUNDS

❖ Take $x \in X = \left\{\text{strings of length } \frac{n}{4} \text{ with weight } d\right\}$

❖ Take $y \in Y = \{y \mid \text{HAM}(x, y) = d \text{ or } \text{HAM}(x, y) = d + 1\}$

❖ Cannot differentiate whether $\text{HAM}(x, y) \leq d$ or $\text{HAM}(x, y) > d$ in $o(d \log n)$ space!

❖ Define $s(x, y) = v^R x y^R v$, where v is the prefix of $10110011100011110000\ldots = 1^1 0^1 1^2 0^2 \ldots$ of length $\frac{n}{4}$ (GMSU16).

YES:
If $\text{HAM}(x, y) \leq d$, then the longest d-near-palindrome is long

NO:
If $\text{HAM}(x, y) > d$, then the longest d-near-palindrome is short

RESULTS (K-PERIODICITY)

❖ $O(k^4 \log^9 n)$ space to find the shortest k-period in one-pass.

❖ $O(k^4 \log^9 n)$ space to find the shortest k-period in two-passes, even if aperiodic.

❖ $\Omega(n)$ space to find the k-period, if aperiodic, in one-pass.

❖ $\Omega(k \log n)$ space to find the k-period, even if periodic, in one-pass.

FULL VERSIONS

❖ [EGSZ17] Funda Ergün, Elena Grigorescu, Erfan Sadeqi Azer, and Samson Zhou. Streaming periodicity with mismatches. RANDOM 2017 (to appear)

❖ [GSZ17] Elena Grigorescu, Erfan Sadeqi Azer, and Samson Zhou. Streaming for Aibohphobes: Longest Palindrome with Mismatches.

THANKS!

FUTURE WORK

❖ What can we say about these problems with other distance metrics (particularly, edit distance)?

❖ Can we improve the space usage? Specifically, the k^4 dependence comes from the structural property, which might have room for improvement.

❖ What if we allow some special characters, such as wild cards?

REFERENCES

❖ [BEMS14] Petra Berenbrink, Funda Ergün, Frederik Mallmann-Trenn, and Erfan Sadeqi Azer. Palindrome recognition in the streaming model. STACS 2014

❖ [CFP + 16] Raphaël Clifford, Allyx Fontaine, Ely Porat, Benjamin Sach, and Tatiana A. Starikovskaya. The k-mismatch problem revisited. SODA 2016

❖ [EJS10] Funda Ergün, Hossein Jowhari, and Mert Saglam. Periodicity in streams. RANDOM 2010

❖ [GMSU16] Pawel Gawrychowski, Oleg Merkurev, Arseny M. Shur, and Przemyslaw Uznanski. Tight tradeoffs for real-time approximation of longest palindromes in streams. CPM 2016