Streaming Periodicity with Mismatches

Funda Ergun,
Elena Grigorescu,
Erfan Sadeqi Azer,
Samson Zhou

Periodicity

- A portion of a string that repeats

  ABCDABCDABCDABCD

  ABCDABCDABCDABCD

Periodicity

- Alternate definition: prefix is the same as suffix
- If S has length n, and $S[1{:}\, n-p] = S[p+1{:}\, n]$, then we say S has period p.

ABCDABCDABCDABCD

ABCDABCDABCD

ABCDABCDABCD

ABCDABCDABCDABCD

Hamming Distance

- Given strings X, Y, the Hamming distance between X and Y is defined as the positions i at which $X_i \neq Y_i$.

S = HAMMING

T = FALLING

$$HAM(S, T) = 3$$

k-Periodicity

- A string that is “almost” periodic, robust to k changes.
- Periodicity: $S[1{:}n-p] = S[p+1{:}n]$
- k-Periodicity: $\mathrm{HAM}(S[1{:}n-p], S[p+1, n]) \leq k$.

  ABCDABCDABCEABCE

  ABCDABCDABCEABCE

  ABCDABCDABCE

  ABCDABCEABCE

  ABCDABCDABCEABCE

  1-period: 4
- Long term periodic changes, but also encompasses “natural” definition.

Streaming Model

- String of length n arrives one symbol at a time
- Use $o(n)$ space, ideally $O(polylog\ n)$

abaacabaccbabbbcbabbccababbccb

abaacabaccbabbbcbabbccababbccb

abaacabaccbabbbcbabbccababbccb

k-Periodicity Problem

- Given a string S of length n, which arrives in a data stream, identify the smallest k-period in space $o(n)$.

- Given a string S of length n, which arrives in a data stream, identify the smallest k-period in space $o(n)$, with two passes.

Related Work

- ❖ $O(\log^2 n)$ space to find the shortest period in one-pass, if $p \leq \frac{n}{2}$. (ErgunJowhariSaglam10)
- ❖ $\Omega(n)$ space to find the period in one-pass, if $p > \frac{n}{2}$. (EJS10)
- ❖ $O(\log^2 n)$ space to find the shortest period in two-passes, even if $p > \frac{n}{2}$. (EJS10)

- ❖ k-Mismatch Problem: $O(k^2 \log^8 n)$ space to find all instances of a pattern P within a text T with up to k errors. (CliffordFontainePoratSachStarikovskaya16)

k-Periodicity (Our results)

- $O(k^4 \log^9 n)$ space to find the shortest k-period in one-pass, if $p \leq \frac{n}{2}$.
- $O(k^4 \log^9 n)$ space to find the shortest k-period in two-passes, even if $p > \frac{n}{2}$.
- $\Omega(n)$ space to find the k-period, if $p > \frac{n}{2}$, in one-pass.
- $\Omega(k \log n)$ space to find the k-period, even if $p \leq \frac{n}{2}$, in one-pass.

Ideas from Streaming Periodicity

- A period p satisfies $S[1: n-p] = S[p+1, n]$.
- If $p \leq \frac{n}{2}$, then $S\left[1: \frac{n}{2}\right] = S\left[p+1, p+\frac{n}{2}\right]$.

  ABCDABCDABCDABCD

  ABCDABCDABCDABCD

  ABCDABCDABCDABCD

  ABCDABCDABCDABCD
- If $p > \frac{n}{2}$, then for some m, $S[1: 2^m] = S[p+1, p+2^m]$.

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)

Ideas from Streaming Periodicity

- First pass: Find all positions p such that first $\frac{n}{2}$ characters match.

$$S\left[1:\frac{n}{2}\right] = S\left[p+1, p+\frac{n}{2}\right].$$

ABCDABCDABCDABCD

ABCDABCDABCDABCD

- Second pass: For each p, check whether p is a k-period.

$$S[1:n-p] = S[p+1,n].$$

ABCDABCDABCDABCD

ABCDABCDABCDABCD

Overall Idea

- A period p satisfies $\mathrm{HAM}(S[1{:}n-p], S[p+1, n]) \le k$.
- If $p \le \frac{n}{2}$, then $\mathrm{HAM}\left(S\left[1{:}\frac{n}{2}\right], S\left[p+1, p+\frac{n}{2}\right]\right) \le k$.
- First pass: Find all positions p that match the first $\frac{n}{2}$ characters.

$$\mathrm{HAM}\left(S\left[1{:}\frac{n}{2}\right], S\left[p+1, p+\frac{n}{2}\right]\right) \le k.$$

- Second pass: For each p, check whether p is a k-period.

$$\mathrm{HAM}(S[1{:}n-p], S[p+1, n]) \le k.$$

- Reduction to Pattern Matching / k-Mismatch

First Pass to Second Pass?

- First pass: Find all positions p, "candidate" k-periods.

$$\mathrm{HAM}\left(S\left[1:\frac{n}{2}\right], S\left[p+1, p+\frac{n}{2}\right]\right) \leq k.$$

- Second pass: For each p, check whether p is a k-period.

$$\mathrm{HAM}(S[1:n-p], S[p+1, n]) \leq k.$$

- ABCDABCDABCDABCDABCD
- Candidate positions $p = \{4,8,12,16, \ldots\}$.
- Candidates form an arithmetic progression!

First Pass to Second Pass?

- If p and q are periods, then $d = \gcd(p, q)$ is a period.
- Does not work for k-periodicity!
- AAAABA, $k = 1$
- $p = 2$: AAAABA, AAAABA

  AAAA
  AABA
  1 mismatch
- $p = 3$: AAAABA, AAAABA

  AAA
  ABA
  1 mismatch
- $p = 1$: AAAABA, AAAABA

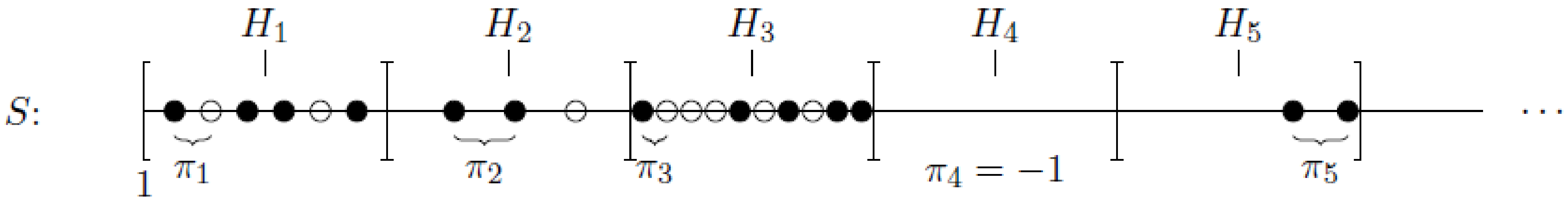
  AAAAB
  AAABA
  2 mismatches!

First Pass to Second Pass?

- Periodicity: Candidate positions $p = \{4,8,12,16,\ldots\}$
  - What's actually happening in the second pass?
  - Using $S[1:4]$, $S[5:8]$, $S[9:12]$,… to build $S[5:n]$, $S[9:n]$, $S[13:n]$,…
  - Can do this because $S[1:4]$, $S[5:8]$, $S[9:12]$ are all the same!
- k-periodicity: Candidate positions $p = \{8,16,20,28,32\ldots\}$?
- Attempt: Candidate positions $p = \{4,8,12,16,20,24,28,32\ldots\}$?
  - Can still do above construction if "most" of $S[1:4]$, $S[5:8]$, $S[9:12]$ are the same
  - Not sure if true…

First Pass to Second Pass?

- Candidates $p = \{8,16,20,27,30,39,45,55\}$?
- Candidates $p = \{8,12,16,20\},\{27,30,33,36,39\},\{45,50,55\}$



Structural Results

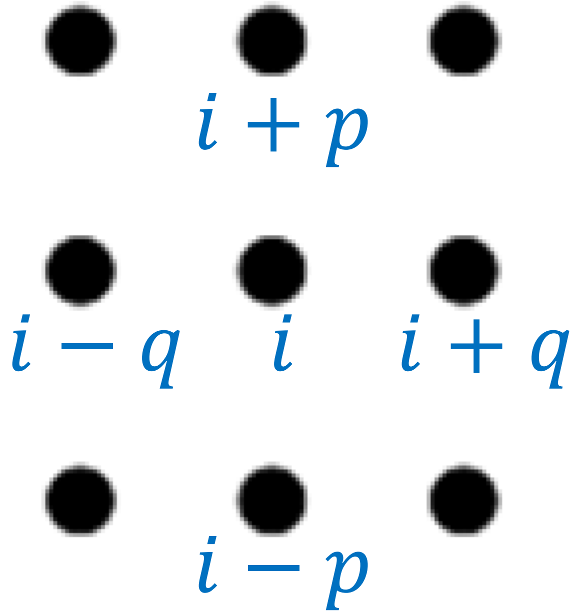
- If p and q are periods, then $d = \gcd(p, q)$ is a period.
- If p and q are "small", then $d = \gcd(p, q)$ is a $O(k^2)$-period.
  - At most $O(k^2)$ of the substrings $S[1:d], S[d+1:2d], S[2d+1:3d]$, can be different

Structural Results

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

If there are at most k indices i such that $S[i] \neq S[i+p]$, and at most k indices j such that $S[j] \neq S[j+q]$, then there are at most $O(k^2)$ indices l such that $S[l] \neq S[l+d]$.

❖ Consider the indices as a grid.

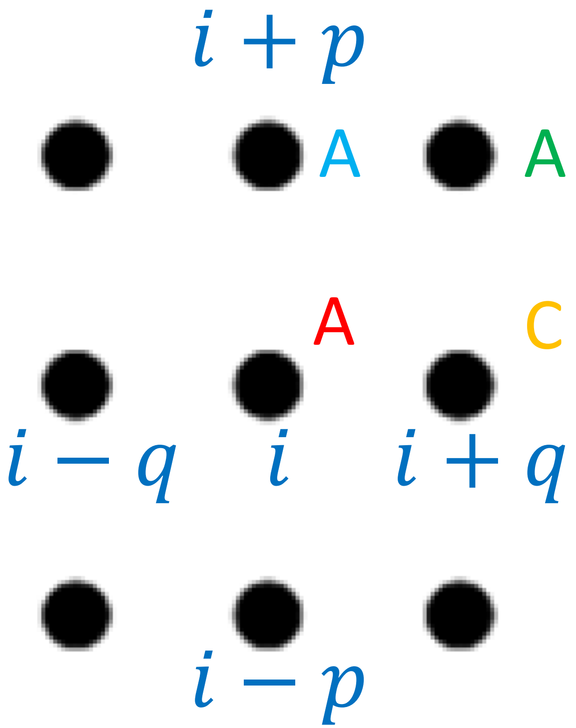


Structural Results

...AABAAABCCAA...

$p = 3, q = 7$

❖ Bound the number of indices l such that $S[l] \neq S[l+d]$.



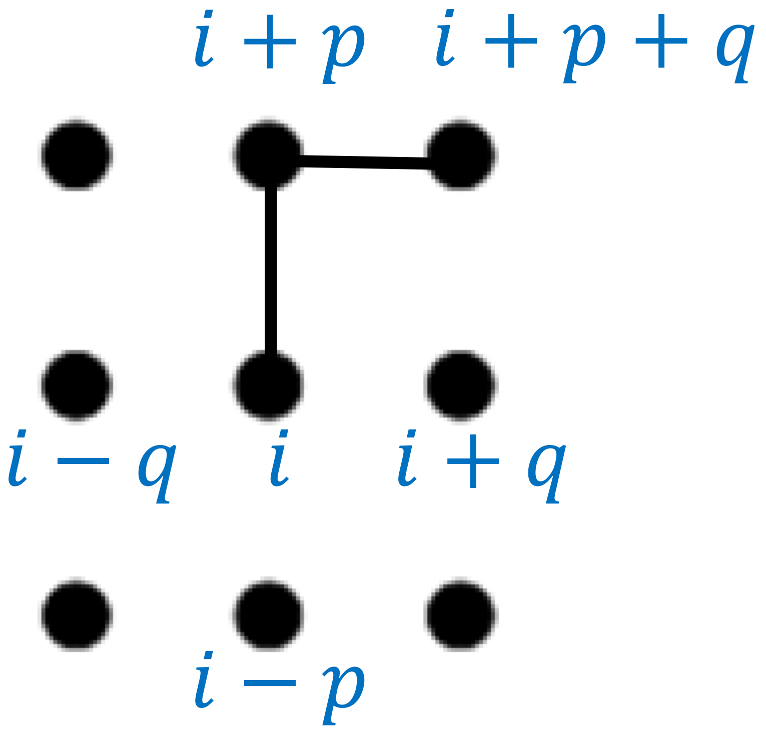
Structural Results

- Connect adjacent points with edges.
- "Good edge" if $S[i] = S[i+p]$.
- "Bad edge" if $S[i] \neq S[i+p]$.
- If there exists a path from i to j which "hops" along good edges, then $S[i] = S[j]$.

…AABAAABCCAA…

$p = 3, q = 7$

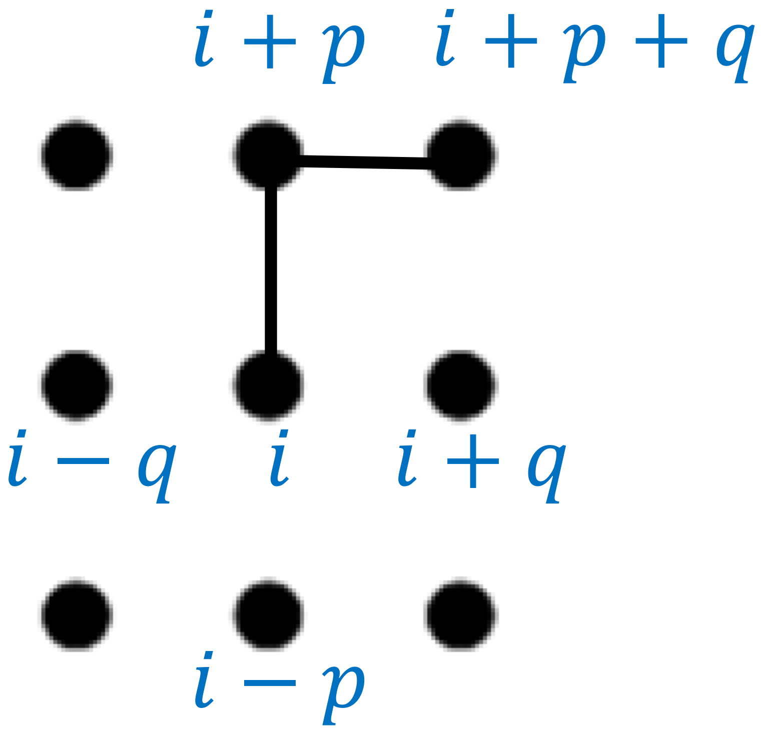
…AABAAABCCAA…



Structural Results

If there are at most k indices i such that $S[i] \neq S[i+p]$, and at most k indices j such that $S[j] \neq S[j+q]$, then there are at most $O(k^2)$ indices l such that $S[l] \neq S[l+d]$.

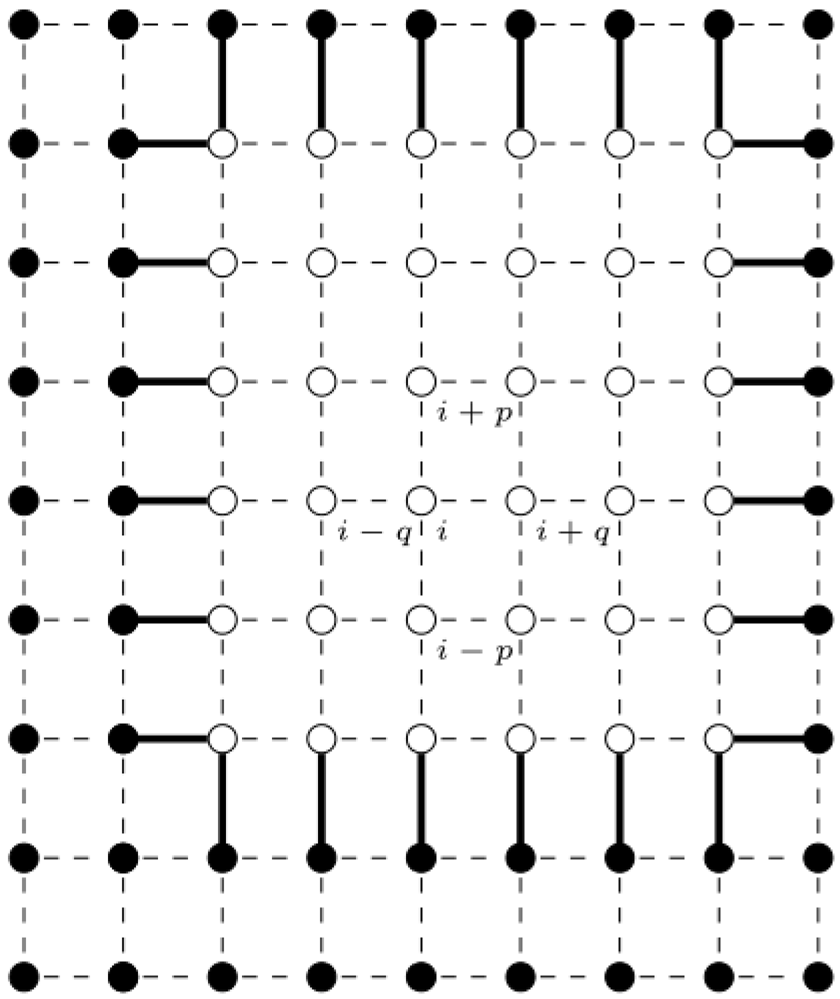
- Bound the number of indices l such that $S[l] \neq S[l+d]$.
- If $S[l] \neq S[l+d]$, then l must be enclosed by bad edges.
- There are at most $2k$ bad edges.
- How many enclosed points can there be?



Structural Results

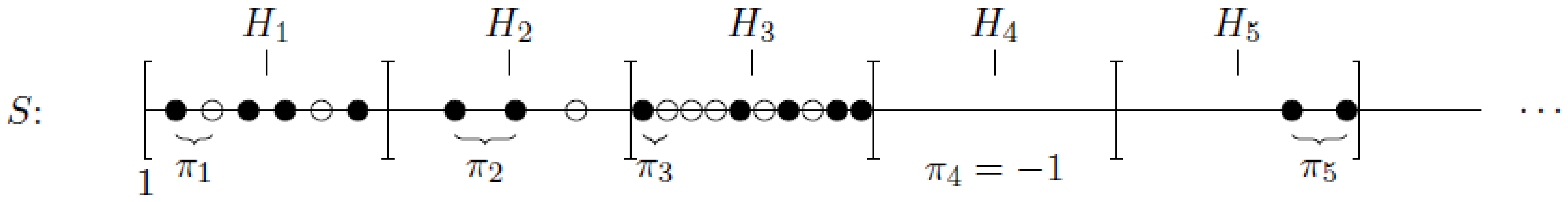
❖ If there are at most $2k$ bad edges, there are $O(k^2)$ enclosed points.

❖ There are $O(k^2)$ indices l such that $S[l] \neq S[l+d]$.



In review

- If p and q are “small”, then $d = \gcd(p, q)$ is a $O(k^2)$-period.
- Positions $p = \{8,16,20,27,30,39,45,55\}$?
- Positions $p = \{8,12,16,20\},\{27,30,33,36,39\},\{45,50,55\}$



In review

- First pass: Find all positions p such that

$$\mathrm{HAM}\left(S\left[1 : \frac{n}{2}\right], S\left[p+1, p+\frac{n}{2}\right]\right) \le k.$$

- Second pass: For each p, check if

$$\mathrm{HAM}(S[1 : n-p], S[p+1, n]) \le k.$$

Open Problems

- 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 and the k-Mismatch Problem algorithm.
- What if we allow some special characters, such as wild cards?

Questions?

thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
bedankt
dziękuje
спасибо
Баярлалаа
рахмат
go raibh maith agat
sukriya
kop khun krap
terima kasih
mochchakkeram
tapadh leat
dank je
grazie
arigatô
dakujem
takk
sagolun
taiku
ευχαριστώ
감사합니다
ありがとう
мерси
хвала
дякую
спасибі
mersi
barka
welalin
tack
spas
kia ora
faafetai lava
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
kiitos
dankie
nanni
nandri
dhanyavad
hvala
maururu
köszönöm
bayarlalaa
gràcie
enkosi
asante
manana
obrigada
murakoze
tenki
djiere dieuf
tau
mamnun
chokrane
akun
dankon
ačiū
chnorakaloutioun
gratias ago
gràcies
sulpáy
sobodi
děkuji
mèsi
didi madloba
najis tuke
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
tanemirt
rahmet
xièxie
diolch
dhanyavadagalu
shukriya
mercé
trugarez