Variance of the Internal Profile in Suffix Trees

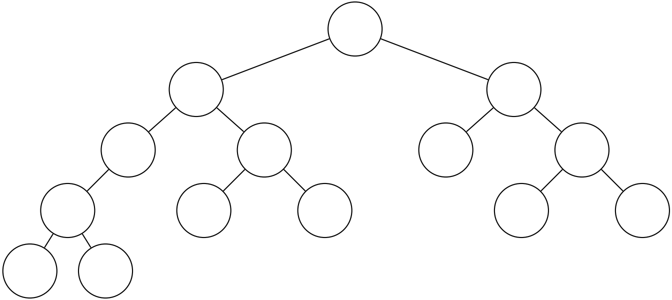
Jeff Gaither[1] and Mark Daniel Ward[2]

[1]Mathematical Biosciences Institute, The Ohio State University
[2]Department of Statistics, Purdue University

Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
July 4, 2016

Quantity of interest – internal profile

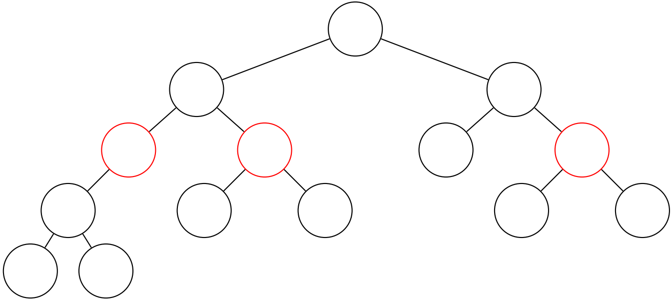
Interested in the internal profile (number of internal nodes) of a suffix tree at level k



Quantity of interest – internal profile

Interested in the internal profile (number of internal nodes) of a suffix tree at level k

In the tree below, there are 3 internal nodes of depth two, so the internal profile at level two is 3.



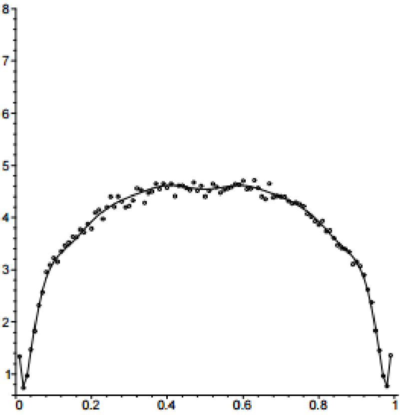
Appeal of problem

We consider variance of internal profile...why is this interesting?

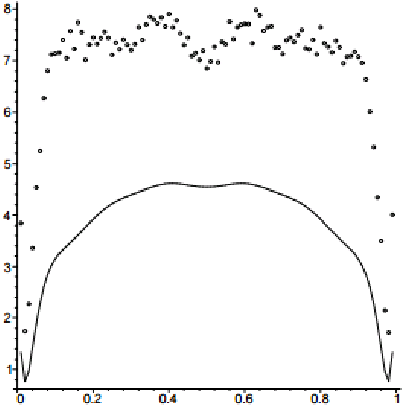
Profile is interesting because it leads to lots of other parameters (most notably total size of tree)

Variance is interesting because it's known to be different in suffix trees and tries

Trie vs. suffix tree - σ of internal profile



Independent model: standard dev.



Dependent model: standard dev.

from *"q-gram analysis and urn models,"* Nicodème, DMTCS 2003

Conjecture

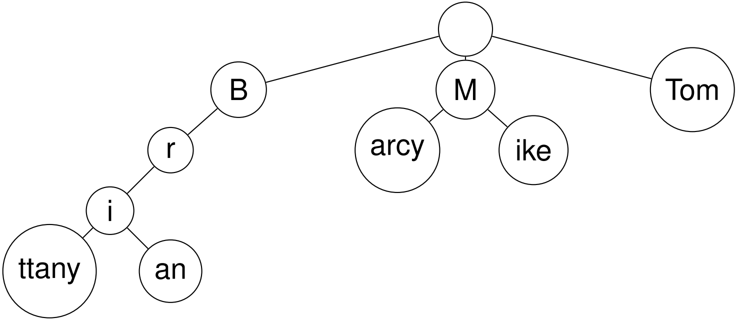
"Regarding variance of a suffix tree, one can derive the generating function... but so far attempts to make it suitable for asymptotic expansion of the variance have not been successful. It is conjectured that the error term between the suffix tree and the independent tries becomes larger than the order of the variance... when the alphabet size is small."
– Jacquet/Szpankowski, *2013*

This is close – in fact, variance in tries and suffix trees has same order, just different coefficients.

Brief review of tries

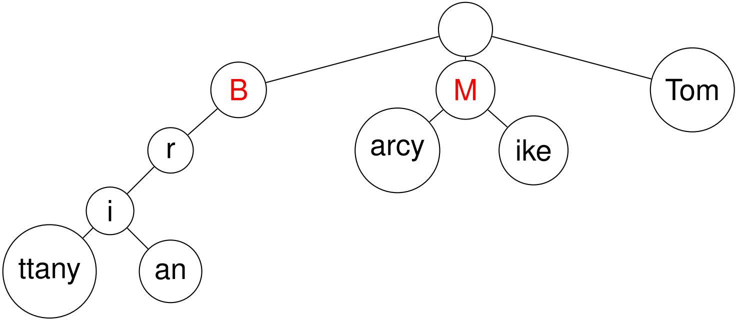
Given collection of independent strings $S_1, S_2, \ldots,$ store each string at node corresponding to shortest distinguishing prefix

Brittany, Brian, Marcy, Mike, Tom



Brief review of tries

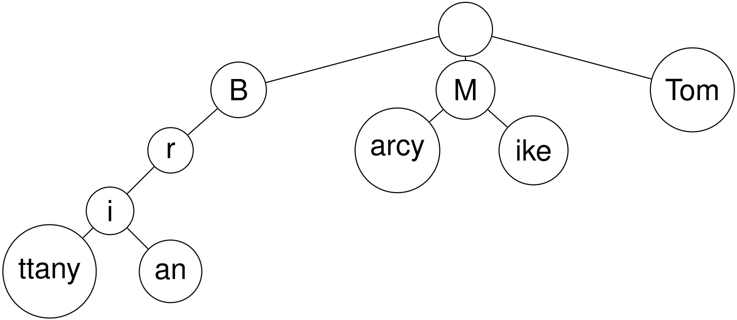
Internal profile at level 1 is 2



Correspondence between nodes and prefixes

Note that the node corresponding to prefix is in profile iff at least two of the generating strings start with that prefix

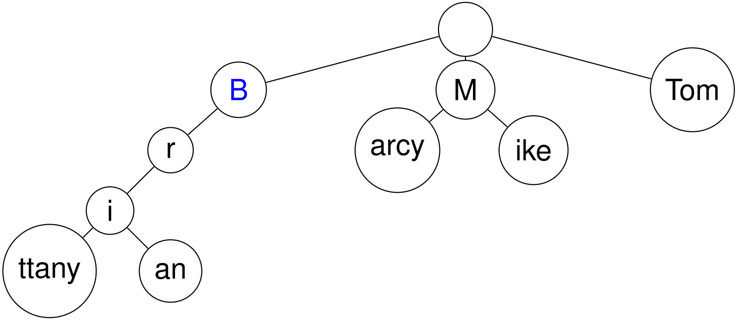
Brittany, Brian, Marcy, Mike, Tom



Brief review of tries

Note that the node corresponding to prefix is in profile iff at least two of the generating strings start with that prefix

Brittany, Brian, Marcy, Mike, Tom

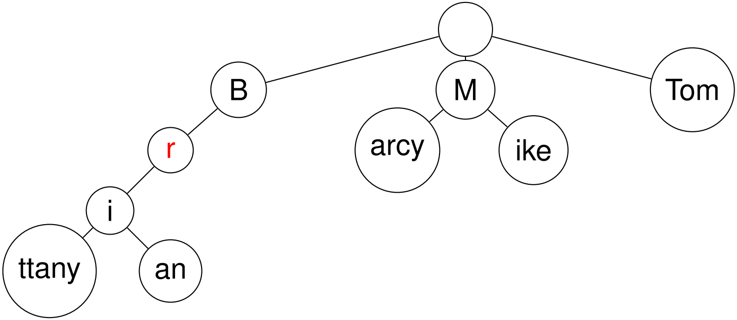


Brief review of tries

Note that the node corresponding to prefix is in profile iff at least two of the generating strings start with that prefix

An example at level $k = 2$:

Brittany, Brian, Marcy, Mike, Tom

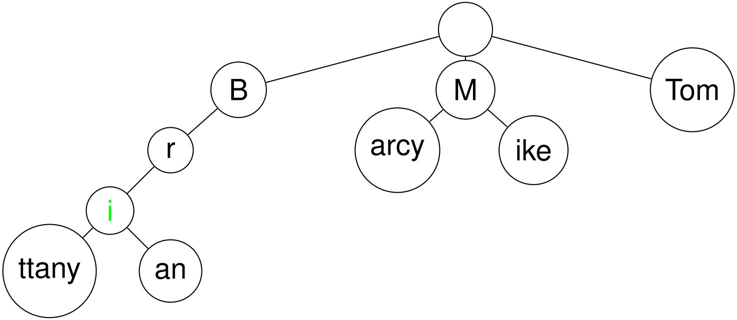


Brief review of tries

Note that the node corresponding to substring contributes to the internal profile iff at least two words begin with that string.

An example at level $k = 3$:

Brittany, Brian, Marcy, Mike, Tom

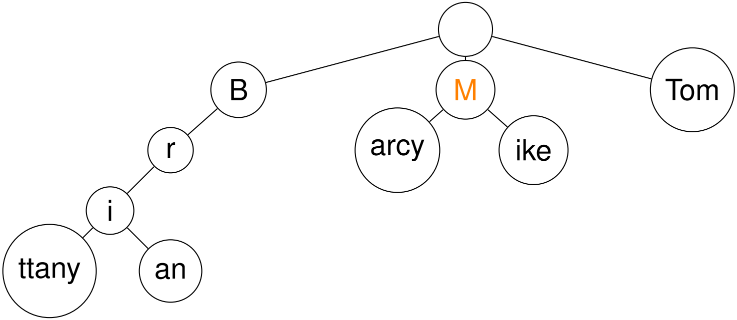


Brief review of tries

Note that the node corresponding to substring contributes to the internal profile iff at least two words begin with that string.

Another example back at level $k = 1$:

Brittany, Brian, Marcy, Mike, Tom



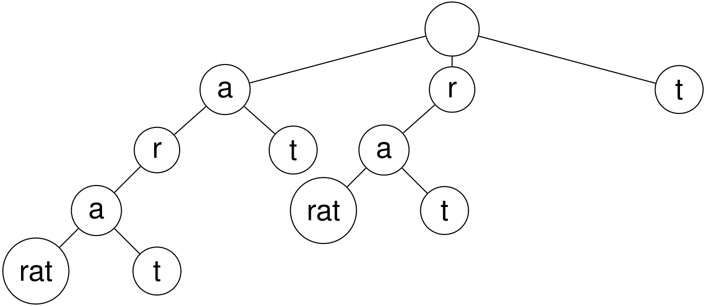
Brief review of suffix trees

Suffix trees are like tries, but built from all suffixes of a single string.

$$S = ararat$$

$$S^{(1)} = ararat \quad S^{(2)} = rarat \quad S^{(3)} = arat$$

$$S^{(4)} = rat \quad S^{(5)} = at \quad S^{(6)} = t$$



Ararat

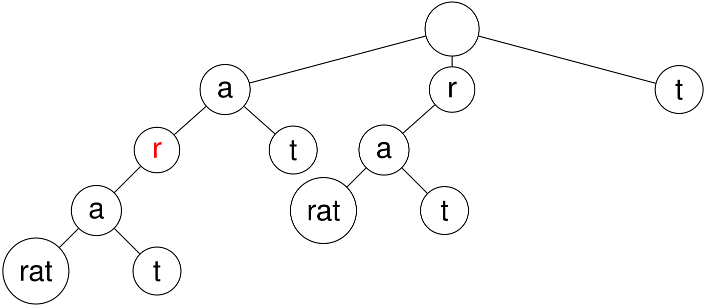
Suffix tree - internal profile

Construction dictates that internal profile is number of substrings of length k that appear at least twice in base-string

$$S = ararat$$

$$S^{(1)} = ararat \quad S^{(2)} = rarat \quad S^{(3)} = arat$$

$$S^{(4)} = rat \quad S^{(5)} = at \quad S^{(6)} = t$$



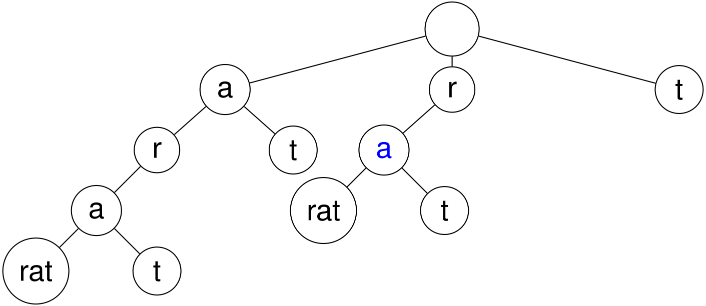
Suffix tree - internal profile

Construction dictates that internal profile is number of substrings of length k that appear at least twice in base-string

$$S = ararat$$

$$S^{(1)} = ararat \quad S^{(2)} = rarat \quad S^{(3)} = arat$$

$$S^{(4)} = rat \quad S^{(5)} = at \quad S^{(6)} = t$$



Strings

Our strings are infinite and built from the binary alphabet $\mathcal{A} = \{a, b\}$.

Given any letter S_i in string S, we have

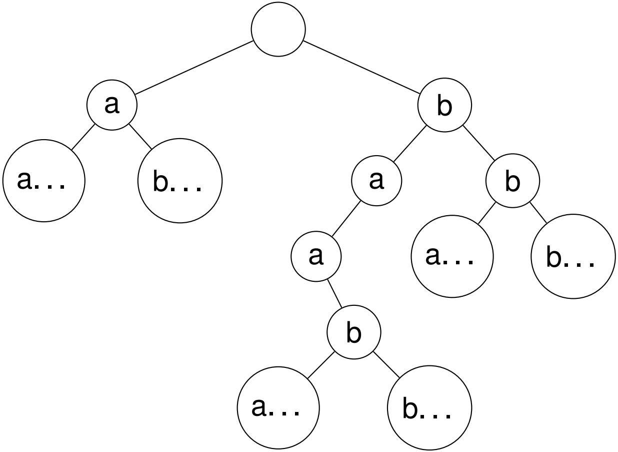
$$\mathbb{P}(S_i = a) = p > \frac{1}{2}; \qquad \mathbb{P}(S_i = b) = q := 1 - p.$$

"Infinite" guarantees that a.s., two suffixes taken from same tree will be distinct (so no unending branches)

Model suffix tree example

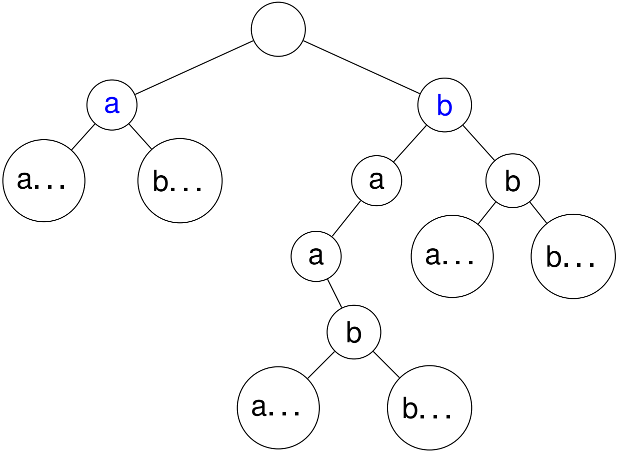
Suffix tree of size 6 built from string

$$S = \color{red}{bbbaabaabbbbabbabb}\ldots$$



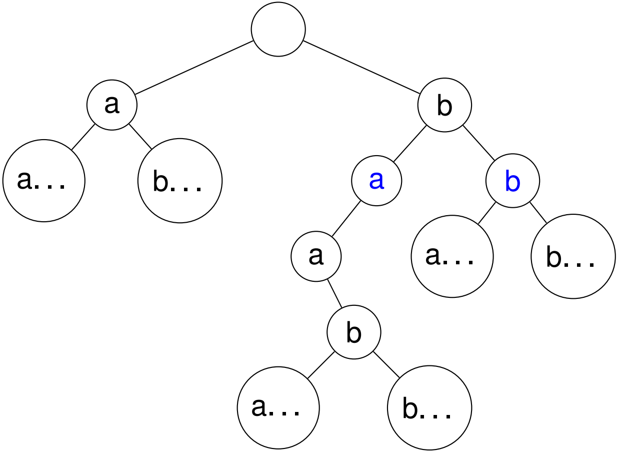
Model suffix tree example

Profile at depth one is 2



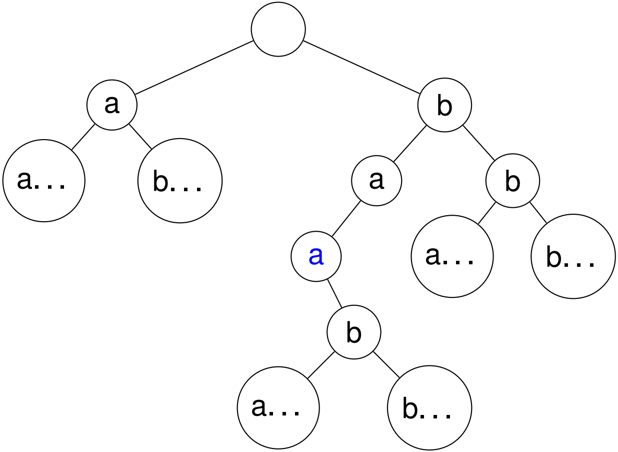
Model suffix tree example

Profile at depth two is 2



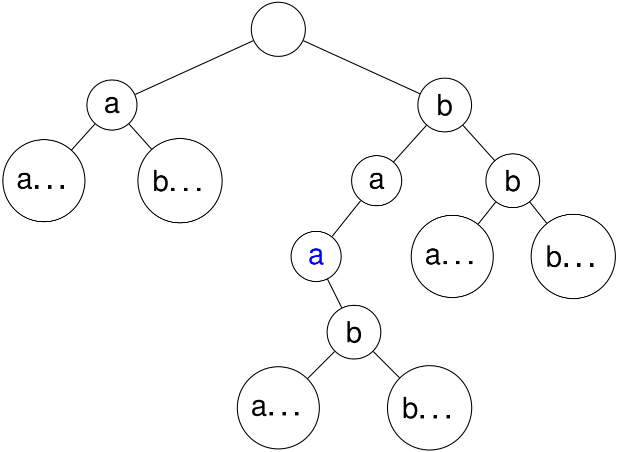
Model suffix tree example

Profile at depth three is 1



Model suffix tree example

Etc...



Scaling depth with size

One last point: what depth k to consider? As number of strings $n \to \infty$, any fixed level k will fill up

Answer: Following Park et al., we assume that

$$\alpha := \lim_{n\to\infty} \frac{k}{\log(n)} \qquad \text{exists.}$$

Let $X_{n,k}$ denote internal profile at level k of suffix tree built from n suffixes, and consider $\mathrm{Var}(X_{n,k})$.

Main results - small alpha

When limit α is small, have easy and very strong bound on the decay of $\mathrm{Var}(X_{n,k})$.

Theorem

When

$$\alpha < \frac{1}{-\log(q)},$$

there exists $B > 0$ *such that*

$$\mathrm{Var}(X_{n,k}) = O(e^{-n^B}).$$

Main result – saddle point regime

Theorem

Define the function

$$h(s) = -s + \alpha \log(p^{-s} + q^{-s})$$

and suppose that

$$\frac{1}{-\log(q)} < \alpha < \frac{p^2 + q^2}{-p^2 \log(p) - q^2 \log(q)}.$$

Then there exists unique $\rho \in (-2, \infty)$ *such that* $h'(\rho) = 0$*, and we have*

$$\mathrm{Var}(X_{n,k}) = \frac{n^{h(\rho)}(C_1(n) + 2C_2(n))}{\sqrt{\log(n)}} \times (1 + O(\log(n)^{-1})).$$

where the $C_i(n)$ *are bounded, positive and with nonzero* lim inf.

Saddle point regime cont'd

Theorem

$$\mathrm{Var}(X_{n,k}) = \frac{n^{h(\rho)}(C_1(n) + 2C_2(n))}{\sqrt{\log(n)}} \times (1 + O(\log(n)^{-1})),$$

The function $C_1(n)$ is given by

$$C_1(n) = \frac{(1 - 2^{-\rho} - \rho 2^{-\rho-2})\Gamma(\rho+2)}{\sqrt{2\pi h''(\rho)}} \times (1 + \textit{small fluctuations})$$

Remark: The $C_1(n)$ portion of our estimate is precisely the variance for a trie, as derived in Park.

Saddle point regime cont'd

The $C_2(n)$ portion of the variance, which is specific to suffix-trees, is built from many different terms whose orders approach order $n^{h(\rho)}$ of trie-term.

$$h(s) = -s + \alpha \log(p^{-s} + q^{-s}),$$

We define extension of $h(s)$,

$$\begin{aligned} H(s, r, c, d) = {} & -s + \alpha(1 - r)\log(p^{-s} + q^{-s}) \\ & - s\left(\frac{\alpha}{k}\right)\log((p^c q^{1-c})^{kr} + (p^d q^{1-d})^{kr}) \end{aligned}$$

with saddle point $\rho_{r,c,d}$.

Theorem

Let $r = \frac{\ell}{k}$, $c = \frac{i}{\ell}$ *and* $d = \frac{j}{\ell}$. *Then* $C_2(n)$ *has the form*

$$C_2(n) = \sum_{\substack{0<\ell<k \\ 0\leq i,j\leq \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} \frac{1}{\sqrt{2\pi \frac{\partial H}{\partial s}(\rho_{r,c,d})}}$$

$$\times \left(\sum_{m\geq 2} \frac{\Gamma(\rho_{r,c,d}+m)}{m!} \Big(\frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}} \Big)^{m-1} \times \Big[(m-1)^2 \right.$$

$$\left. \frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}} + m(2-m) + m(\rho_{r,c,d}+m) \frac{p^i q^{\ell-i} p^j q^{\ell-j}}{(p^i q^{\ell-i} + p^j q^{\ell-j})^2} \Big] \right)$$

$$\times (1 + \textit{small fluctuations}).$$

Also, we have $\sum_{\ell\geq \ell_0} \cdots = O(n^{-\beta(\ell_0/k)})$ *for a* $\beta > 0$
(the sum is concentrated near $\ell = 1$.*)*

Polar regime

Behavior when $\alpha > \frac{p^2+q^2}{-p^2\log(p)-q^2\log(q)}$ is pretty much the same, except we use $s = -2$ in place of ρ, and a lot of terms disappear.

Polar regime

Behavior when $\alpha > \frac{p^2+q^2}{-p^2\log(p)-q^2\log(q)}$ is pretty much the same, except we use $s = -2$ in place of ρ, and a lot of terms disappear.

Theorem

Suppose $\alpha > \alpha_2 = \frac{p^2+q^2}{-p^2\log(p)-q^2\log(q)}$. *Then for some* $\epsilon > 0$, *we have*

$$\mathrm{Var}(X_{n,k}) = n^{h(-2)} \times (\tilde{C}_1(n) + 2\tilde{C}_2(n)) \times \left(1 + O(n^{-\epsilon})\right),$$

where

$$\tilde{C}_1(n) = 1$$

$$\tilde{C}2(n) = \sum_{\substack{0<\ell<k \\ 0\le i,j\le \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(-2,r,c,d)}}{n^{h(-2)}} \times \left(\frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}}\right).$$

Considering $C_2(n)$

We note two things about new suffix-tree variance coefficient $C_2(n)$.

Considering $C_2(n)$

We note two things about new suffix-tree variance coefficient $C_2(n)$.

1. It's not very neat

Considering $C_2(n)$

We note two things about new suffix-tree variance coefficient $C_2(n)$.
1. It's not very neat

$$C_2(n) = \sum_{\substack{0<\ell<k \\ 0\le i,j\le \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} \frac{1}{\sqrt{2\pi \frac{\partial H}{\partial s}(\rho_{r,c,d})}}$$

$$\times \left(\sum_{m\ge 2} \frac{\Gamma(\rho_{r,c,d}+m)}{m!} \left(\frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}} \right)^{m-1} \times \left[(m-1)^2 \right.\right.$$

$$\left.\left. \frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}} + m(2-m) + m(\rho_{r,c,d}+m) q \frac{p^i q^{\ell-i} p^j q^{\ell-j}}{(p^i q^{\ell-i} + p^j q^{\ell-j})^2} \right] \right)$$

$$\times (1 + \text{small fluctuations}).$$

with $r = \frac{\ell}{k}$, $c = \frac{i}{\ell}$ and $d = \frac{j}{\ell}$.

Considering $C_2(n)$

Even if we compress it, it's still got these binomial coefficient and this weird quotient of powers of n.

$$C_2(n) = \sum_{\substack{0<\ell<k \\ 0\leq i,j\leq \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} \frac{1}{\sqrt{2\pi \frac{\partial H}{\partial s}(\rho_{r,c,d})}}$$

$$\times \left(\sum_{m\geq 2} \frac{\Gamma(\rho_{r,c,d}+m)}{m!} \left(\frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}}\right)^{m-1} \times \left[(m-1)^2 \right.\right.$$

$$\left.\left. \frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}} + m(2-m) + m(\rho_{r,c,d}+m) q \frac{p^i q^{\ell-i} p^j q^{\ell-j}}{(p^i q^{\ell-i} + p^j q^{\ell-j})^2} \right] \right)$$

$$\times (1 + \text{small fluctuations}).$$

with $r = \frac{\ell}{k}$, $c = \frac{i}{\ell}$ and $d = \frac{j}{\ell}$.

Considering $C_2(n)$

Even if we compress it, it's still got these binomial coefficients and this quotient of powers of n.

$$C_2(n) = \sum_{\substack{0<\ell<k \\ 0\le i,j\le \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} \times W(r,c,d)$$

with $r = \frac{\ell}{k}$, $c = \frac{i}{\ell}$ and $d = \frac{j}{\ell}$.

Considering $C_2(n)$

Even if we compress it, it's still got these binomial coefficients and this quotient of powers of n.

$$C_2(n) = \sum_{\substack{0<\ell<k \\ 0\leq i,j\leq \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} \times W(r,c,d)$$

with $r = \frac{\ell}{k}$, $c = \frac{i}{\ell}$ and $d = \frac{j}{\ell}$.

Not a very good asymptotic coefficient. How do we even know it converges? How do we know it doesn't blow up? Can't we get a closed form?

Mathematical viability

Can show that

$$\binom{\ell}{i}\binom{\ell}{j}\frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} < 1$$

by showing that the map

$$r \to \binom{kr}{krc}\binom{kr}{krd} n^{H(\rho_{r,c,d},r,c,d)}$$

is decreasing in r.

Mathematical viability

And decay condition, that partial sum beyond any ℓ_0 is

$$\sum_{\substack{\ell_0 \leq \ell < k \\ 0 \leq i,j \leq \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} W(n,r,c,d) = O(n^{-(\ell_0/k)\beta})$$

is actually quite strong.

Implies that we can sum ℓ to $\log(\log(k))$ (or something even smaller) rather than k,

since

$$n^{-(\log(\log(k))/k)\beta} = e^{-\log(n)(\log(\log(k))/k)\beta)} = e^{-\log(\log(k))/\alpha}.$$

Aesthetics

But head of sum *does* contribute, so it must be included, pretty or no

The proof sheds some light on form of sum

Quick sketch of proof

Let I_u indicate that node corresponding to word u (of length k) appears at least 2 times (within the first $n+k-1$ characters)
Then profile can be written

$$X_{n,k} = \sum_{u \in \mathcal{A}^k} I_{n,u},$$

and

$$\begin{aligned}\mathrm{Var}(X_{n,k}) &= \sum_{u \in \mathcal{A}^k} \mathrm{Var}(I_{n,u}) \\ &+ \sum_{\substack{u,v \in \mathcal{A}^k \\ u \neq v}} \mathrm{Cov}(I_{n,u}, I_{n,v})\end{aligned}$$

Quick sketch of proof

Let I_u indicate that node corresponding to word u (of length k) appears at least 2 times (within the first $n+k-1$ characters)
Then profile can be written

$$X_{n,k} = \sum_{u \in \mathcal{A}^k} I_{n,u},$$

and

$$\begin{aligned} \mathrm{Var}(X_{n,k}) &= \sum_{u \in \mathcal{A}^k} \mathrm{Var}(I_{n,u}) && \text{trie term} \\ &+ \sum_{\substack{u,v \in \mathcal{A}^k \\ u \neq v}} \mathrm{Cov}(I_{n,u}, I_{n,v}) && \text{new suffix-tree term} \end{aligned}$$

Correlations significant

After analysis, covariances $\mathrm{Cov}(I_{n,u}, Iv)$, $u \neq v$ turn out to contain terms like

$$n^2\mathbb{P}(u)\mathbb{P}(v)e^{-n(\mathbb{P}(u)+\mathbb{P}(v))}(e^{n(\mathbb{P}(u)C_{u,v}(1)+\mathbb{P}(v)C_{v,u}(1))} - 1)$$

This term

$$\mathbb{P}(u)C_{u,v}(1) + \mathbb{P}(v)C_{v,u}(1)$$

appears to be a genuine novelty

Correlation polynomials

Terms $C_{u,v}(1)$ and $C_{v,u}(1)$ are *correlation polynomials*

They measure the degree of overlap between *a* and *b*.

Always really small, unless (rarely!) there is a really long suffix of u that is also a prefix of v

$$u = baaaaaaa, \quad v = aaaaaaaa$$

A classic lemma by Jacquet and Szpankowski states that this is vanishingly improbable when $u = v$, except for the trivial complete self-overlap...

$$\sum_{u \in \mathcal{A}^k} \mathbb{P}(u)|C_{u,v}(1) - 1| = O(p^{k/2})$$

Simultaneous overlaps are negligible

We would therefore expect terms $\mathbb{P}(u)C_{u,v}(1)$ to be negligible in some formalizable sense.

In fact, term $\mathbb{P}(u)C_{u,v}(1)C_{v,u}(1)$ IS negligible,

$$\sum_{u,v} \mathbb{P}(u)C_{u,v}(1)C_{v,u}(1) = O(p^{k/2})$$

which is to say we can ignore the possibility that $C_{u,v}(1)$ and $C_{v,u}(1)$ will *simultaneously* be large...

The meaning of H

However, when use the approximation

$$\binom{kr}{krc}\binom{kr}{krd}n^{H(\rho_{r,c,d},r,c,d)}$$

for the contribution u, v pairs whose overlapping proportion is $1-r$, and whose nonshared regions contain the respective proportions c and d of a's,

we find that this is maximized when $r=0$, i.e. when the overlap is total.

Hence, all word-pairs with exceptional overlap contribute. Effect tapers off quickly, but there's no sharp dividing line where we can say "the asymototic contribution ends here"

Inevitability of sum

Thus, we must sum over at least the germ of the sum...

but as slowly divergent a germ as we like.

$$C_2(n) \approx \sum_{\substack{0<\ell<\log(\log(\log(\log(k)))) \\ 0\le i,j\le \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} W(n,r,c,d)$$

Acknowledgements

Thanks to Mark Daniel Ward

and to support from

Mathematical Biosciences Institute

Conclusion

Thank you for your attention!