Heather C. Smith

Research Statement

School of Mathematics
Georgia Institute of Technology
heather.smith@math.gatech.edu
http://people.math.gatech.edu/~hsmith90

As a discrete mathematician, two themes in my research are (1) discrete mathematical biology and (2) the combinatorics of partially ordered sets. However, I have a broad range of interests. Computational complexity, crossing numbers, saturation, posets, biological modelsI have some experience in each of these areas and a wide variety of open problems to explore further. As I highly value collaboration, I am eager to expand my knowledge and share my skills in working with new colleagues. I am excited to continue my mathematical research and contribute to the life of a vibrant department.

Many biological processes, such as genome rearrangement, naturally lend themselves to discrete models wherein mathematical tools can be used for analysis. On the other hand, these models often motivate new and interesting mathematics. My work spans both ends of the spectrum.

For genome rearrangement, a common mode of molecular evolution, biologists need fast algorithms for near-uniform sampling of evolutionary histories for hypothesis testing. We study the computational complexity of discrete models for gene rearrangement.

To work more closely with biological data and learn first-hand which problems are most relevant for biologists, I pursued interdisciplinary research, collaborating with Professor Jordan's bioinformatics lab at Georgia Tech. Together, we designed a fast, scalable algorithm for bacterial typing which is critical for epidemiological surveillance and outbreak control.

Biological structures also motivate new mathematics. A tanglegram is a particular type of graph whose crossing number is related to the number of times parasites switch hosts and to the number of horizontal gene transfers. This connection initiated our work on extremal problems regarding the crossing numbers of tanglegrams.

The combinatorics of partially ordered sets (posets) has a long and growing history. Dimension, in particular, has many important connections to other graph parameters. We are interested in two variants of poset dimension, one only a year old, comparing their behavior with dimension, relative to other graph parameters. Probabilistic techniques and Ramsey theoretic tools have been key in our work.

Extremal problems for forbidden posets have traditionally sought the largest family of sets that avoid a specific configuration. We take a new perspective on these extremal problems, introducing a minimization problem that parallels the notion of graph saturation. We prove a logarithmic lower bound on the saturation number for a large family of posets and give exact values for some small forbidden posets.

1. Discrete Mathematical Biology

1.1. **Computational Complexity of Genome Rearrangement.** Genome rearrangement was first identified in the pioneering work of Sturtevant and Dobzhansky [31, 10] only a few years after the first genetic map [30]. With the prevalence of mutations in the genome, biologists have a huge amount information for the analysis of evolutionary relationships among genes and among taxa. In the age of genomics, when whole genome sequencing became technically possible and affordable, gene order and gene content data emerged as a new source of information. Only a small fraction of the available data has been analyzed, notwithstanding efforts for automation and increasing computing power.

We need biologically relevant models for genome rearrangement which are computationally approachable, since the sheer number of evolutionary histories makes it infeasible to check statistics on every possibility. We seek methods for sampling histories from a near-uniform distribution for hypothesis testing. If a polynomial time algorithm exists for enumerating histories under a given model, then the histories can be sampled from a near-uniform distribution. Even if the problem is #P-complete (the analogue of NP-completeness for enumeration problems), there may still be a polynomial time algorithm for near-uniform sampling in the form of a rapidly mixing Markov chain on the space of all evolutionary histories.

My dissertation focused on the Single Cut-or-Join model for genome rearrangement. In this model, genomes are represented by edge-labeled directed graphs with maximum total degree two. Each edge corresponds to a gene and the direction identifies the DNA strand which contains the coding sequence for that gene. Each component of the underlying graph corresponds to a chromosome where paths are linear chromosomes, as in the genomes of eukaryotes, and cycles are circular chromosomes, which play a role in tumor growth [26].

In the digraph, vertices with total degree one are *telomeres* while vertices with total degree two are *adjacencies*. A genome with a given set of genes is uniquely identifiable by its adjacencies. Genome rearrangement is a modification of the telomeres and adjacencies while the nucleotide sequences of genes remain unchanged.

The Single Cut-or-Join model [14] for genome rearrangement has two allowable moves:

Cut (Fission): Replace an adjacency with two telomeres.
Join (Fusion): Replace two telomeres with an adjacency.

This model gives first approximations for evolutionary distances very quickly. It was also the first model with a polynomial time algorithm for the most parsimonious median problem: Given a family of genomes $\mathcal{G}$, enumerate the genomes G_m (called medians) such that $\sum_{G\in\mathcal{G}} d_{\text{SCJ}}(G_m, G)$ is minimum.

A most parsimonious median scenario for a family of observed genomes $\mathcal{G}$ is a tuple $(G_m, \{\sigma_G\}_{G\in\mathcal{G}})$ where G_m is a median and σ_G a minimum length series of cut and joins that transform G_m into G. We proved the following for enumerating these histories:

Theorem 1.1. [24] *For the Single Cut-or-Join model, the most parsimonious median scenario problem is #P-complete.*

Despite the fact that an exact enumeration is infeasible, it is important to ask the following:

Problem 1.2. *Is there a polynomial time near-uniform sampler for the most parsimonious median scenario problem?*

We defined a standard Markov chain on the space of most parsimonious median scenarios and proved that it was not rapidly mixing due to a bottleneck found using boosting (blow up) techniques. However, this does not exclude the possibility of a sampler. In fact, our proof for Theorem 1.1 uses modulo prime calculations, a technique which has been utilized for problems that have a rapidly mixing Markov chain. A solution to the following problem could imply a solution for Problem 1.2:

Problem 1.3. *Given a family of genomes $\mathcal{G}$, what is the complexity of locating the median which appears in the maximum number of most parsimonious median scenarios.*

If Problem 1.3 is NP-hard, then the answer to Problem 1.2 is negative under the assumption $RP \neq NP$. On the other hand, a constructive polynomial time algorithm for Problem 1.3 could lend the necessary insight to design a more sophisticated Markov chain for Problem 1.2.

I recently submitted an NSF grant proposal to continue research on these problems for the Single Cut-or-Join model and related complexity problems for other genome rearrangement models. As a survey of these models, Miklós and I have a paper in *BMC Bioinformatics* [23] which chronicles the state-of-the-art knowledge for the computational complexity of the following three models: Reversals, Double Cut-and-Join, and Single Cut-or-Join.

Related work:

I. Miklós and H. Smith. "Sampling and counting genome rearrangement scenarios," *BMC Bioinformatics* 16(Suppl 14):S6 (2015). (*DOI:10.1186/1471-2105-16-S14-S6*, 2015).
I. Miklós and H. Smith. "The computational complexity of calculating partition functions of optimal medians with Hamming distance," *Submitted, 2016* Preprint at arXiv:1506.06107v2 [cs.CC].

1.2. **Bacterial Typing.** Having worked with genome rearrangement models for my dissertation, I was eager to begin an interdisciplinary collaboration during my postdoc to move my research closer to the applications. In meetings with Professor Jordan's bioinformatics lab, I offered mathematical tools for some of the problems they were researching.

Together, we designed a fast, scalable algorithm for bacterial typing which is critical for epidemiological surveillance and outbreak control. This assembly-free and alignment-free tool takes the raw sequence reads as input and scales efficiently to whole genome typing schemes.

Bacterial genomes are sorted into types (or strains). Traditionally, types have been assigned based on the DNA sequences corresponding to a commonly accepted collection of housekeeping genes. In order to determine the type of a specific sample, the genome is first sequenced, producing a large collection of raw sequence reads, each approximately 100 base pairs in length, such that each base pair in the genome is covered approximately 30 times. At this point, these raw reads must be assembled to obtain the genome sequence, a process in which de Bruijn graphs are often useful. Then alignment occurs in a search for the housekeeping genes. At this point, the appropriate DNA sequences can be compared to a database to determine the sequence type of the bacteria.

Our tool avoids assembly and alignment, working directly from the raw sequence reads to determine the sequence type. We are not only able to determine the type based on the multilocus sequence typing scheme (MLST), but our tool also extends to whole genome typing, a typing scheme that is becoming more useful with the growing amounts of genome data.

Related work:

H. F. Espitia, A. T. Chande, H. Smith, I. K. Jordan, and L. Rishishwar. "STing: ultrafast sequence typing with in silico aptamers," *In preparation.*

1.3. **Tanglegrams and Crossing Numbers.** Tanglegrams were well-studied in the phylogenetics literature before graph theorists became interested in the tanglegram crossing number, a quantity that has been related to the number of times parasites switch hosts [19] and the number of horizontal gene transfers [6]. A tanglegram of order n consists of an ordered pair of binary trees, each with n leaves, with a matching between their leaves. A layout of a tanglegram is a plane drawing of the trees, the first drawn in the half plane $x \leq 0$ with its leaves on the line $x = 0$ and the second in the half plane $x \geq 1$ with its leaves on the line $x = 1$, with a straight line drawing of the matching edges between the leaves. The crossing number of a layout is the number of pairs of matching edges that cross and the tanglegram crossing number, denoted $\mathrm{crt}(T)$, is the minimum crossing number over all layouts.

Since computing the tanglegram crossing number is NP-hard [15], it is important to find fast algorithms for approximation. Given a tanglegram T, partition the leaves of each binary tree B_1 and B_2 into large *clades*, sets of leaves that remain in a consecutive block in any layout of T. Let $w_{i,j}$ be the number of edges in the matching with one endpoint in clade U_i from B_1 and one endpoint in clade V_j from B_2. For any pair of clades $\{U_a, U_b\}$ from B_1 and any clades $\{V_c, V_d\}$ in B_2, the contribution to the tangle crossing number of T is at least $\min\{w_{ua}w_{vb}, w_{ub}w_{va}\}$. Through simulations on tanglegrams chosen uniformly at random, we found that tgus nubunyn is on average $0.055n^2 + O(n)$, which agrees with the expected tanglegram crossing number of $\Theta(n^2)$ [8]. To explain the output of the simulations, we are working on the following problem where $\mathcal{T}_n$ is the set of tanglegrams of order n.

Problem 1.4. *For a tanglegram selected uniformly at random from $\mathcal{T}_n$, what is the expected lower bound on the tanglegram crossing number obtained from this algorithm?*

Paralleling results for the crossing number, we answer two extremal questions.

Theorem 1.5. [1] *For any $T \in \mathcal{T}_n$, $\mathrm{crt}(T) < \frac{1}{2}\binom{n}{2}$. Further, $\lim_{n\to\infty} \max_{T\in\mathcal{T}_n} \mathrm{crt}(T)/\binom{n}{2}$ exists and equals $\frac{1}{2}$.*

Theorem 1.6. [1] *For any $T \in \mathcal{T}_n$ and any matching edge e in T, let T' be the sub-tanglegram that results when e is removed and the corresponding leaves are deleted. Then $\mathrm{crt}(T) - \mathrm{crt}(T') \leq n - 3$. This inequality is best possible, even when T' is planar.*

Related work:

R. Anderson, S. Bai, F. Barrera-Cruz, É. Czabarka, G. Da Lozzo, N. L. F. Hobson, J. C.-H. Lin, A. Mohr, H. C. Smith, L. Székely, and H. Whitlatch. "Analogies between the crossing number and the tangle crossing number," *Submitted* Preprint at arXiv:1709.08119 [math.CO].

2. Combinatorics of Partially Ordered Sets

2.1. **Variants of dimension.** Poset dimension, defined by Dushnik and Miller [11], has many beautiful connections with other graph parameters. For example, a graph is planar if and only if the dimension of its incidence poset is at most 3 (Schnyder's Theorem [27]).

In 2016, Uekerdt introduced *local dimension*, a new concept which sparked renewed interest in Boolean dimension [17]. It is important to compare the behavior of these quantities, determining which theorems for dimension extend to local dimension and Boolean dimension. At least half a dozen papers on these topics have been written in the past 6 months. Our paper [3] makes new connections between local dimension and structural graph theory in addition to extending many inequalities that hold for dimension to local dimension.

A *realizer* of a poset P is a non-empty set of linear extensions $\mathcal{L}$ such that if two points x and y in P are incomparable, then $x < y$ in some $L \in \mathcal{L}$ and $x > y$ in some $L' \in \mathcal{L}$. The *dimension* is the least d for which there is a realizer of size d. The standard example, S_n, on $2n$ points with minimal points $\{a_i : i \in [n]\}$ and maximal points $\{b_i : i \in [n]\}$ has $a_i < b_j$ if and only if $i \neq j$. This poset has dimension n because each linear extension can only realize $a_i > b_i$ for one i. This is the maximum dimension realized by a poset with $2n$ points (Hiraguchi [21]).

A *local realizer* of a poset P is a non-empty set of partial linear extensions $\mathcal{L}$, linear extensions of induced subposets of P, such that the following two properties hold:

* If $x < y$ in P, then $x < y$ in some $L \in \mathcal{L}$.
* If x and y are incomparable in P, then $x < y$ in some $L \in \mathcal{L}$ and $x > y$ in some $L' \in \mathcal{L}$.

The *local dimension* is the least d for which there is a local realizer in which each element in P appears at most d times. Because a linear extension is a partial linear extension, the local dimension of a poset is at most its dimension. Further, the local dimension of the standard example is only 3 as witnessed by 2 linear extensions

and n partial linear extensions of the form $b_i < a_i$. Despite this, local dimension is an unbounded quantity. For example, Ramsey-theoretic tools show that for any d and large enough n, the poset $P(1,2;n)$, containing a point for each subset of size 1 or 2 and ordered by inclusion, has dimension d.

While the maximum dimension of a poset with n points is $n/2$, a relatively easy result to prove, this bound is not tight for local dimension. We use a result of Erdős and Pyber [13] for covering graphs with complete bipartite graphs to show that the maximum local dimension of a poset with n points is $O(n/\log n)$. We proved that this is tight:

Theorem 2.1. [22] *The maximum local dimension of poset with n points is $\Omega\left(\frac{n}{\log n}\right)$.*

We use probabilistic methods to prove the lower bound, extending a result of Chung, Erdős, and Spencer [7] to difference graphs, bipartite graphs with nested neighborhoods, which correspond with partial linear extensions. The analogous maximum of $\Theta(\log n)$ for Boolean dimension is tight via a simple counting argument.

The standard example, described above, is a subposet of the planar Kelly poset whose cover graph has path-width three. Thus dimension is not bounded by a function of the path-width of the cover graph. However, there is a function $d(t,h)$ which bounds the dimension of a height h poset whose cover graph has tree-width t.

Since the local dimension of the standard example, and the Kelly poset, is three, it is plausible that local dimension could be bounded in terms of the path-width of the cover graph. The following pair of results reveals the behavior for local dimension:

Theorem 2.2. [3]

- *If P is a poset whose cover graph has path-width t, then the local dimension of P is $O(5^{(t+1)^2})$.*
- *For any d, there is a poset P whose cover graph has tree-width 3 and local dimension at least d.*

The second result uses Ramsey-theoretic tools for binary trees, a tool that proved useful in another context. It is well-known that the clique number, $\omega(G)$, and the chromatic number, $\chi(G)$, are separated by shift graphs. Seymour [28] introduced the tree-chromatic number, tree-$\chi(G)$, and the path-chromatic number, path-$\chi(G)$, for which $\omega(G) \leq \text{tree-}\chi(G) \leq \text{path-}\chi(G) \leq \chi(G)$. The chromatic number of a tree-decomposition is the maximum chromatic number of a subgraph induced by the vertices in a single bag. The tree-chromatic number is the minimum chromatic number of a tree-decomposition. Path-chromatic number is defined similarly. Shift graphs separate the chromatic number and the path-chromatic number while the classic Erdős construction [12] for graphs with large chromatic number and large girth have large tree-chromatic number [28]. We separated the remaining pair with the following result:

Theorem 2.3. [2] *For each n, there is a graph H_n with tree-chromatic number 2 and path-chromatic number at least n.*

Here is a selection of open problems surrounding local dimension and Boolean dimension.

Problem 2.4.

- *If P has large local (Boolean) dimension, which structures must be present as subposets of P?*
- *There is a family of planar posets P_n such that the local dimension of P_n is at least n* [5]*. Does such a family exist for Boolean dimension?*
- *What is the maximum local (Boolean) dimension of a poset with width w?*
- *What is the local (Boolean) dimension of the subset lattice?*

Related work:

F. Barrera-Cruz, S. Felsner, T. Mészáros, P. Micek, H. Smith, L. Taylor, and W. T. Trotter. "Separating tree-chromatic number from path-chromatic number," *Submitted* Preprint at arXiv:1703.03973 [math.CO].
F. Barrera-Cruz, T. Prag, H. Smith, L. Taylor, W. T. Trotter. "Comparing Dusknik-Miller dimension, Boolean dimension, and local dimension," *Submitted* Preprint at arXiv:1710.09467 [math.CO].
J. Kim, R. R. Martin, T. Masařík, W. Shull, H. Smith, A. Uzzell, and Z. Wang. "Local dimension and size," *In preparation.*

2.2. **Saturation for posets.** Among extremal problems for forbidden subposets, the quantity $\text{La}(n,P)$ is the maximum number of points in the subset lattice $\mathcal{B}_n$ that does not contain P as a (weak) subposet. We introduce a minimization problem for forbidden subposets which parallels saturation for graphs.

Sperner's Theorem [29], one of the first in this area, gives $\text{La}(n,\mathcal{C}_2) = \binom{n}{\lfloor n/2 \rfloor}$, the maximum number of points which avoid a chain of height 2, also the maximum size of an antichain. For the diamond, $\mathcal{B}_2$, the smallest poset for which $\text{La}(n,P)$ is unknown, Griggs and Lu conjecture $\text{La}(n,\mathcal{B}_2) \leq (2+o(1))\binom{n}{\lfloor n/2 \rfloor}$.

If we merely ask for the minimum number of points that avoid a subposet P, the answer is trivially zero. However, if we add the constraint that any strict superset must contain P as an induced subposet, we have an induced-P-saturated family, a notion which parallels graph saturation. Let $\text{sat}^*(n, P)$ be the minimum number of points in an induced-P-saturated family in $\mathcal{B}_n$. The non-induced version of poset saturation was introduced in [18], but it was only examined for chains where the two definitions align.

If $\mathcal{C}_k$ is the chain with k points and $\mathcal{A}_k$ is the antichain with k points, one can quickly see that $\text{sat}^*(n, \mathcal{C}_2) = 1$, $\text{sat}^*(n, \mathcal{C}_3) = 2$, $\text{sat}^*(n, \mathcal{A}_2) = n+1$, and $\text{sat}^*(n, \mathcal{A}_3) = 2n$, the latter two making use of Dilworth's Theorem [9]. For the three point poset $\mathcal{V}_2$ which has a unique minimum element and two maximum elements, we proved $\text{sat}^*(n, \mathcal{V}_2) = n + 1$ for $n \geq 2$.

Using families of sets that separate points and biclique cover numbers, we proved the following lower bound:

Theorem 2.5. [16] *Let P be a poset with at least two points and the following property: If an element $x \in P$ has precisely one cover $y \in P$, then y covers at least two elements in P. Then $\lceil \log_2 n \rceil \leq \text{sat}^*(n, P)$.*

For the diamond, $\text{sat}^*(n, \mathcal{B}_2)$ is between $\lceil \log_2 n \rceil$ and $n + 1$, but the exact value remains unknown. We suspect the right answer is $n + 1$ since we proved any $\mathcal{B}_2$-saturating family which contains $\emptyset$ or $[n]$ has size at least $n + 1$.

Related work:

M. Ferrara, B. Kay, L. Kramer, R. R. Martin, B. Reiniger, H. Smith, and E. Sullivan. "The saturation number of induced subposets of the Boolean lattice," *Discrete Mathematics* 340.10 (2017): 2479-2487.

3. Student Research

Many students have shared my interest in problems motivated by biology. I first studied a combinatorial model for RNA with graduate students at a summer workshop, and later explored a variation with an undergraduate. In 2016, I co-led an REU at Georgia Tech where I was the lead on a research project involving Markov chains on meanders, working closely with two undergraduates and a graduate assistant. These students have since given talks in two undergraduate research seminars and a poster at the 2017 Joint Mathematics Meetings.

3.1. **Combinatorial Model for RNA-folding.** The molecule ribonucleic acid (RNA) consists of a single strand of nucleotides, representable by a finite word from the alphabet A, U, C, and G. In contrast to the double helix of DNA, the single-stranded RNA often folds onto itself as the nucleotides form bonds. Predicting the folded structure of RNA is important as the folded structure indicates its functionality.

Black, Drellich, and Tymoczko [4] introduced a generalized combinatorial model considering words from the alphabet $\{A_1, \overline{A}_1, A_2, \overline{A}_2, \ldots, A_m, \overline{A}_m\}$ where A_i and $\overline{A}_i$ bond like the Watson-Crick pairs. A word w is foldable if can be written around a rooted plane tree T, starting at the root and working counterclockwise so that the two letters labeling the same edge are complements. The tree T is called w-valid.

We define a bijection between edge-colored rooted plane trees and words folded onto trees which is used to enumerate words for which there is only one valid tree. In addition, we examine the set $\mathcal{R}(n, m)$ consisting of all integers k for which there exists a word of length $2n$ with exactly k w-valid trees. We determine a large consecutive sequence of values in $\mathcal{R}(n, 1)$ and also find intervals which are missing from $\mathcal{R}(n, 1)$.

My undergraduate student gave partial results on the following problem.

Problem 3.1. *Consider a word w on alphabet $\{A, U, C, G\}$. If we allow for the bonds AU, CG, and GU, give a fast algorithm to determine if w is foldable.*

Related work:

B. Bjorkman, G. Cochran, W. Gao, L. Keough, R. Kirsch, M. Phillipson, D. Rorabaugh, H. Smith, and J. Wise. "k-foldability of words," *Submitted* Preprint at arXiv: 1710.10616 [math.CO].

3.2. **Markov chains on Meanders.** A meander of order n is a non-self-intersecting single closed curve in the plane which crosses a horizontal line in $2n$ points. While the study of meanders dates back to Poincaré [25], exact enumeration of meanders remains an open problem. In an effort to estimate the number of meanders via Markov chain sampling methods, we define a local move on meanders and study properties of the resulting state space graph $\mathcal{G}$. While our local move is a restriction of the move studied by Heitsch and Tetali [20], their techiques do not carry over to our setting.

Before we can analyzing mixing times, we must prove that the Markov chain on $\mathcal{G}$ is irreducible. Specifically, is $\mathcal{G}$ connected? Given an affirmatice answer, the next step is to analyze mixing time. During the REU, we proved a partial result for connectedness, considering the class of meanders obtained by Kreweras complementation. We proved the following:

Theorem 3.2 (De Vierno, Nang, Prag, Smith)**.** *The subgraph of $\mathcal{G}$ induced on the set of meanders which are Kreweras complements is connected.*

References

[1] R. Anderson, S. Bai, F. Barrera-Cruz, É. Czabarka, G. Da Lozzo, N. L. F. Hobson, J. C.-H. Lin, A. Mohr, H. Smith, L. Székely, and H. Whitlatch. Analogies between the crossing number and the tangle crossing number. *Submitted*, 2017. Available at arXiv:1709.08119[math.CO].

[2] F. Barrera-Cruz, S. Felsner, T. Mészáros, P. Micek, H. Smith, L. Taylor, and W. T. Trotter. Separating tree-chromatic number from path-chromatic number. *Submitted*, 2017. Available at arXiv:1703.03973v2 [math.CO].

[3] F. Barrera-Cruz, T. Prag, H. Smith, L. Taylor, and W. T. Trotter. Comparing Dushnik-Miller dimension, Boolean dimension and local dimension. *Submitted*, 2017.

[4] F. Black, E. Drellich, and J. Tymoczko. Valid plane trees: Combinatorial models for RNA secondary structures with Watson-Crick base pairs. *To appear*, 2017. Available at arXiv:1501.03238v2[math.CO].

[5] B. Bosek, H. Grytczuk, and W. T. Trotter. Local dimension is unbounded for planar posets. *preliminary manuscript*, 2017.

[6] A. Burt and R. Trivers. *Genes in Conflict: The Biology of Selfish Genetic Elements*. Belknap Press, Cambridge, MA, 2008.

[7] F. R. K. Chung, P. Erdős, and J. Spencer. On the decomposition of graphs into complete bipartite subgraphs. In P. Erdős, L. Alpár, G. Halász, and A. Sárközy, editors, *Studies in Pure Mathematics*, pages 95–101, 1983.

[8] É. Czabarka, L. A. Székely, and S. Wagner. Inducibility in binary trees and crossings in random tanglegrams. *SIAM J. Disc. Math.*, 31(3):1732–1750, 2017.

[9] R. P. Dilworth. A decomposition theorem for partially ordered sets. *Annals of Mathematics*, 51(1):161–166, 1950.

[10] Th. Dobzhansky and A. H. Sturtevant. Inversions in the chromosomes of Drosophila Pseudoobscura. *Genetics*, 23(1):28–64, 1938.

[11] B. Dushnik and E. W. Miller. Partially ordered sets. *American Journal of Mathematics*, 63:600–610, 1941.

[12] P. Erdős. Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38, 1959.

[13] P. Erdős and L. Pyber. Covering a graph by complete bipartite graphs. *Discrete Mathematics*, 170:249–251, 1997.

[14] P. Feijão and J. Meidanis. Scj: A novel rearrangement operation for which sorting, genome median and genome halving problems are easy. *Lecture Notes in Bioinformatics*, 5724:85–96, 2009.

[15] H. Fernau, M. Kaufmann, and M. Poths. Comparing trees via crossing minimization. In S. Sarukkai and S. Sen, editors, *Proc. 25th Intern. Conf. Found. Softw. Techn. Theoret. Comput. Sci. (FSTTCS'05), LNCS vol. 3821*, pages 457–469. Springer-Verlag, 2005.

[16] M. Ferrara, B. Kay, L. Kramer, R. R. Martin, B. Reiniger, H. Smith, and E. Sullivan. The saturation number of induced subposets of the Boolean lattice. *Discrete Mathematics*, 340(10):2479–2487, 2017.

[17] G. Gambosi, J. Nešetřil, and M. Talamo. Boolean representations and quick path searching. In *Automata, Languages and Programming, 14th International Colloquium, ICALP87, Proceedings, Lecture Note Series in Computer Science*, volume 267, pages 404–424, 1987.

[18] D. Gerbner, B. Keszegh, N. Lemons, C. Palmer, D. Pálvölgyi, and B. Patkós. Saturating Sperner families. *Graphs and Combinatorics*, 29(5):1355–1364, 2013.

[19] M. S. Hafner and S. A. Nadler. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, 332:258–259, 1988.

[20] C. Heitsch and P. Tetali. Meander graphs. *Discrete Mathematics and Theoretical Computer Science Proceedings (FPSAC 2011)*, AO, 2011.

[21] T. Hiraguchi. On the dimension of orders. *Sci. Rep. Kanazawa Univ.*, 4:1–20, 1955.

[22] J. Kim, R. R. Martin, T. Masařík, W. Shull, H. Smith, A. Uzzell, and Z. Wang. On the decomposition of graphs into difference graphs and the local dimension of posets. *In preparation.*

[23] I. Miklós and H. Smith. Sampling and counting genome rearrangement scenarios. *BMC Bioinformatics*, 16((Suppl 14):S6), 2015.

[24] I. Miklós and H. Smith. The computational complexity of calculating partition functions of medians with Hamming distance. *Submitted*, 2016. Available at arXiv:1506.06107v2 [cs.CC].

[25] H. Poincaré. Sur un théorème de géométrie. *Rendiconti del Circolo Matematico di Palermo*, 33:375–407, 1912.

[26] B. Raphael and P. Pevzner. Reconstructing tumor amplisomes. *Bioinformatics*, 20(Suppl 1):i265–i273, 2004. (Special ISMB/ECCB 2004 Issue).

[27] W. Schnyder. Planar graphs and poset dimension. *Order*, 5:323–343, 1989.

[28] P. Seymour. Tree-chromatic number. *Journal of Combinatorial Theory, Series B*, 116:229–237, 2016.

[29] E. Sperner. Ein satz uber utermegen einer endlichen menge. *Mathematische Zeitschrift*, 27:544–548, 1928.

[30] A. H. Sturtevant. The linear arrangement of six sex-linked factoris in *Drosophila*, as hown by their mode of association. *Journal of Experimental Zoology*, 14:43–59, 1913.

[31] A. H. Sturtevant and Th. Dobzhansky. Inversions in the third chomosome of wild races of Drosophila pseudoobscura, and their use in the study of the history of the species. *Proceedings of the Nationall Academy of Sciences*, 22(7):448–450, 1936.