Heather C. Smith's Research Statement

My research lies in combinatorics and graph theory. While I am a pure mathematician by choice, I am attracted to problems that further the research in mathematics as well as in other scientific fields. Fortunately, there is a continuous flow of problems and ideas between discrete mathematics and scientific applications. In fact, the National Science Foundation recognized that the biggest obstacle in the progress of life sciences is the lack of mathematical, algorithmic, and statistical tools to handle the amount of data they generate. This concern has accelerated their support of developments in the mathematical sciences.

In 2010, I earned a master's degree in mathematics at Virginia Commonwealth University under the guidance of Richard Hammack with a project on the direct product of digraphs. After this, I started a Ph.D. in mathematics at the University of South Carolina (USC) with research advisor László Székely with projects on finite set systems and extremal problems on trees. At the end of my first year, I received the department award for "2012/2013 Outstanding First Year Graduate Student."

In 2013, I received the SPARC Fellowship from the Office of the Vice President for Research at USC for which the application was a grant proposal. In the writing process, I learned about the balance required to describe the project to a broad audience while providing enough rigor to communicate the difficulty and my qualifications to tackle it.

The following year, I was awarded the Dean's Doctoral Dissertation Fellowship from the College of Arts and Sciences at USC. These two fellowships allowed me to travel to the Alfréd Rényi Institute of Mathematics of the Hungarian Academy of Sciences in Budapest for two month-long research visits to collaborate with István Miklós, who has a Ph.D. in mathematics as well as a master's degree in biology. He is interested in building a rigorous mathematical foundation for bioinformatics and computational biology. In a project on the complexity of genome rearrangement problems, I learned the challenges that come with transferring biological ideas into mathematical models, capturing necessary details while simplifying other parameters.

Last summer, I attended two different workshops as a fully-funded participant. At the Mathematics Research Community (MRC) on Network Science, sponsored by the AMS, I began a network modeling project with an applied mathematician and two statisticians. This project was continued during a trip to London and Oxford supported by the AMS. In our diversity, I learned skills for communication and directing research to satisfy everyone's interests.

I also attended the Rocky Mountain–Great Plains Graduate Research Workshop in Combinatorics (GRWC) which was co-hosted by the University of Colorado Denver and the University of Denver. We began a project on saturation in posets. The combining of graph saturation and poset theory emphasized the strength in collaboration. At the workshop, I was also introduced to Sage and have since used it for other projects.

I have had the opportunity to collaborate with researchers in mathematics, computer science, statistics, and biology. To make use of existing mathematical results and techniques, many of the problems were translated from different fields into mathematical models. The outcome has been a variety of projects in discrete mathematics that are also the source of new problems. My results and current projects fall under the categories of computational complexity, finite set systems and posets, extremal problems on trees, graph games, graph products, and network modeling. I will describe them in more detail.

1. Computational Complexity

[Collaboration with István Miklós, [15]]

With the discrepancy between P and NP, there are some problems for which we cannot hope to compute exact answers in a reasonable amount of time. In an effort to continue exploring such problems, scientists are willing to accept good approximations. As a result, computational complexity has expanded beyond complexity classes P and NP.

Decision problems have analogous counting problems. For example, rather than asking if a Hamilton cycle exists in a graph, we may ask how many Hamilton cycles exist. The class #P consists of those counting problems which correspond to the decision problems in NP. Like NP-complete, the class #P-complete consists of those problems in #P which are among the hardest in the class. For problems in #P, we may want to obtain good approximations. A fully polynomial randomized approximation scheme (FPRAS) is a randomized algorithm which runs in polynomial time and will output, with high probability, an approximation that is within a small factor of the correct answer. We proved that the median problem for the Single Cut-or-Join model lies in #P, and in some cases has an FPRAS. Before I describe this problem mathematically, I will first give the biological motivation.

Biologists use trees (and in some cases more complicated networks) to represent the ancestral history of taxonomic units, for examples, species and cancerous cells. Similar models are used in stemmatology to study the history of old manuscripts. The leaves of the tree represent the current generation and edges represent ancestral relations.

Nodes are labeled with strings which may stand for genomic or protein sequences, collections of characteristics, or any other information whose change we are concerned with in our model.

In most applications, only the strings for the leaves of the presumed tree are known. The goal is to infer the most likely tree structure and corresponding ancestral strings according to a reasonable criterion. One of the simplest criterion to evaluate the likelihood is *parsimony*, a popular model among biologists which follows Occam's Razor. Given a tree and strings for its nodes, the parsimony score is the sum of the distances between strings on adjacent pairs of nodes. Distance may be defined in various ways, the simplest one being the Hamming distance. For any tree, a most parsimonious assignment is the one with smallest parsimony score. A tree and labeling is most parsimonious if it has the least parsimony score over all possible trees and sequence assignments. The goal is to find a tree with a most parsimonious labeling.

Due to the double stranded nature of DNA, we represent genes with directed edges. *Genomes* can then be viewed as edge-labelled oriented graphs where the components of the underlying undirected graph are cycles and paths. Those vertices of total degree two are called *adjacencies* while those with total degree one are called *telomeres.* A genome can be characterized by its list of adjacencies. We use this characterization to encode genomes with a binary string.

To describe the change in gene order, we use the simplistic *Single Cut-or-Join* model. This model allows for mutations of genetic structures by sequences of "cuts" (fission) which change a single adjacency into two telomeres and "joins" (fusion) which create an adjacency from two telomeres. One genome can be transformed into another through a sequence of these two simple operations. These sequences are called *sorting scenarios.*

It is easy to show that the Hamming distance between the binary string labels is precisely the number of cuts and joins required to transform one genome into the next. We further simplify our setting by considering only adjacencies which are independent. In other words, the necessary cuts and joins can be performed in any order. Hence, the number of possible sorting scenarios between two genomes is exactly the factorial of the Hamming distance. By selecting a sorting scenario for each edge, we obtain a *tree scenario.* The number of tree scenarios is the product of the number of sorting scenarios for each edge.

Assume now that we have a fixed tree T. Let $\mathcal{L}$ be the collection of most parsimonious string assignments ℓ on T. To gain more information about $\mathcal{L}$, we define a real-valued function f on its elements and ask about the complexity of calculating $\sum_{\ell\in\mathcal{L}} f(\ell)$. For example, when $f(\ell) = 1$ for all $\ell \in \mathcal{L}$, we are asking for the complexity of calculating the size of $\mathcal{L}$. Erdős and Székely [9] state that this can be easily calculated.

Using $\triangle$ for the Hamming distance, let $f(\ell) = \prod_{uv\in E(T)}(\ell(u)\triangle\ell(v))!$ which is precisely the number of tree scenarios for ℓ. We proved that if T is a binary tree or a star, then it is #P-complete to compute $\sum_{\ell\in\mathcal{L}} f(\ell)$. The natural next question has to do with approximability. For binary trees, this quantity cannot be approximated with an FPRAS unless RP=NP [14]. On the other hand, when T is a star, we proved that the most natural Markov chain on $\mathcal{L}$ is torpidly mixing, so it cannot be used for an FPRAS. It is still unknown whether or not there is an FPRAS for the star.

Next, we defined $f(\ell) = \prod_{uv\in E(T)} h(\ell(u)\triangle\ell(v))$ where h is a real-valued function. Depending on the concavity of the logarithm of $h(x)$, we obtained several more general results on the complexity of computing $\sum_{\ell\in\mathcal{L}} f(\ell)$.

Lastly, let $f(\ell)$ count the number of edges in T on which at least one mutation occurs under the labeling ℓ. When T is binary, we showed that it is NP-hard to find the minimum or maximum value of $f(\ell)$.

While our work focused on the Single Cut-or-Join model for genome rearrangement, there are other models which are biologically more accurate for describing changes in gene order. Proving that one of these models admits an FPRAS will result in a polynomial time algorithm to generate random samplings of scenarios under a near-uniform distribution. This opens the way to test various hypotheses about the past.

2. Finite Set Systems and Posets

2.1. **Constructing Baranyai Partitions.** [Collaboration with László Székely, [17]]

Baranyai's theorem states: "For positive integers k, n such that k divides n, the edges of a complete, k-uniform hypergraph on n vertices can be partitioned into $\binom{n-1}{k-1}$ families, each of which is a perfect matching of the vertices." This theorem, which requires only one line to imply the Erdős-Ko-Rado Theorem for $k|n$, was open for 118 years before Baranyai [2] gave a proof in 1975 making use of network flows. However, his proof sheds little light on the structure of the partitions. There is a well-known, circular configuration which yields a straightforward construction for $k = 2$. Beth [5] gave an algebraic construction for $k = 3$. However, neither method has a known extension for larger k.

We have explored bijections between k-ary rooted trees and partitions. From one particular bijection, we developed a recursive construction to build classes of labeled binary trees that encode Baranyai partitions for $k = 2$. Further, the result is provably different from any that can be obtained using the well-known circular construction. It is our hope that a similar bijection with labeled trees or other combinatorial objects will lead to a more general construction of Baranyai partitions. Indeed any construction for $k > 3$ is likely to have an impact and move the area forward.

2.2. **Saturation in Posets.** [Collaboration with Sarah Behrens, Ed Boehnlein, Michael Ferrara, Bill Kay, Lucas Kramer, Luke Nelsen, Ben Reiniger, Derrick Stolee, and Eric Sullivan started at the GRWC]

Given two posets $\mathcal{A} = (A, \leq_A)$ and $\mathcal{B} = (B, \leq_B)$, we say $\mathcal{A}$ is a subposet of $\mathcal{B}$ provided there is an injective map $\varphi : A \hookrightarrow B$ in which $\varphi(x) \leq_B \varphi(y)$ if $x \leq_A y$. If additionally $\varphi(x) \leq_B \varphi(y)$ only if $x \leq_A y$, then we say that $\mathcal{A}$ is an induced subposet of $\mathcal{B}$. For posets $\mathcal{A}$ and $\mathcal{B}$, an induced subposet $\mathcal{F} = (F, \leq_F)$ of $\mathcal{B}$ is (induced) $\mathcal{A}$-saturated if (1) it does not contain $\mathcal{A}$ as a subposet and (2) every induced subposet $\mathcal{F}' = (F', \leq_{F'})$ of $\mathcal{B}$ with $F \subsetneq F'$ has $\mathcal{A}$ as a(n) (induced) subposet.

Fixing $\mathcal{A}$ and $\mathcal{B}$, Ferrara asked for the minimum size of an induced subposet of $\mathcal{B}$ which is (induced) $\mathcal{A}$-saturated. We began our work with $\mathcal{B} = \mathcal{B}_n$, the Boolean lattice on n elements, and considered a variety of basic posets for $\mathcal{A}$ including chains, anti-chains, and diamonds. Specifically, the diamond that we consider is $\mathcal{D}_4$ consisting of four distinct sets, A, B, C, D with $A \subset B \subset D$ and $A \subset C \subset D$ with B and C incomparable. We have established a number of non-trivial upper bounds on these saturation numbers. However, the lower bounds appear to require additional machinery. For $\mathcal{D}_4$, we proved that a minimum size induced $\mathcal{D}_4$-saturated poset has at most $n+1$ elements and every $\mathcal{D}_4$-saturated poset which contains either $\emptyset$ or $[n]$ must have at least $n+1$ elements. We continue to explore lower bounds for $\mathcal{D}_4$.

In addition to completing the lower bound for $\mathcal{D}_4$, a wide variety of open problems exists as we vary $\mathcal{A}$ and $\mathcal{B}$. For example, I would like to consider other small posets including more general diamonds, crowns, or even small Boolean lattices for $\mathcal{A}$. We can also expand our sights to different posets $\mathcal{B}$, such as the divisor lattice or the partition lattice.

3. Extremal Problems on Trees

One problem in the study of social networks involves locating the most important or most influential vertex. There are many vertex measures that highlight various characteristics, such as the degree of a node or number of shortest paths passing through a node. The following two projects give results toward three different centrality measures: eccentricity, distance, and number of subtrees.

3.1. **Eccentricity in Trees.** [Collaboration with László Székely and Hua Wang, [19]]

The eccentricity of a vertex, $ecc_T(v) = max\{d(u, v) : u \in V(T)\}$, was one of the first, distance-based, tree invariants studied [13]. The total eccentricity of a tree, $Ecc(T)$, is the sum of eccentricities of its vertices. It is easy to see that the star minimizes $Ecc(T)$ among trees with a given order, and Dankelmann, Goddard, and Swart [7] showed that the path maximizes $Ecc(T)$. We turn our attention to the more delicate ratios $Ecc(T)/ecc(u)$, $Ecc(T)/ecc(v)$, $ecc(u)/ecc(v)$, and $ecc(u)/ecc(w)$ where u, w are leaves of T and v is in the center of T. For each, we provide extremal values as well as characterize extremal tree structures. Analogous problems have been resolved for other tree invariants including distance [3] and number of subtrees [21].

For a further look at total eccentricity, we examine the class of trees with given degree sequence $(d_1, d_2, \ldots, d_n)$ where $d_i \geq d_{i+1}$ and $d_k > d_{k+1} = 1$ for some $1 \leq k < n$. One of these graphs is the greedy caterpillar [22]. To build one, start with a path P on k vertices which will be the caterpillar's spine. In order to realize the degree sequence, add pendant vertices to each vertex on the spine such that $deg(v) \leq deg(u)$ if v is closer than u to the center of the spine. Another tree in this class is the greedy tree [23]. This is a rooted tree with the largest degree appearing at the root, r, and for any two vertices $u, v \in V(T)$, if $d(u, r) \geq d(v, r)$ then $deg(u) \leq deg(v)$. Further, the tree can be drawn in the plane with vertices at distance i from the root drawn on the line $y = i$ and, whenever u is to the left of v on the same horizontal line, then $deg(u) \geq deg(v)$. We prove that the greedy caterpillar maximizes $Ecc(T)$ and the greedy tree minimizes $Ecc(T)$ among trees with the same degree sequence.

3.2. **Middle Parts of Trees.** [Collaboration with László Székely, Hua Wang, and Shuai Yuan, [18]]

In this project, we broaden our sights to three different vertex attributes for trees, namely the eccentricity of a vertex, the distance of a vertex, and the number of subtrees containing a vertex. The distance of a vertex v is defined to be $\sum_{u \in V(T)} d(u, v)$. Each of these attributes has a naturally defined "middle" part: center, centroid, and subtree core, respectively. We examined the interaction between these middle parts.

There are many trees, such as the star, in which all three middle parts coincide. We proved upper bounds for the distance between pairs of middle parts and constructed families of trees which realized the bounds. Because the distance between middle parts is correlated with the maximum degree and the diameter of the graph, we fixed these quantities, one at a time, to examine the restricted classes. For each of these, we again determined the maximum distance between pairs of middle parts and exhibited corresponding extremal trees.

We further examined the number of subtrees attribute to search among the rooted trees of a given order and height for ones which minimize the number of subtrees containing the root. We proved a list of necessary properties about the desired tree. However, a full characterization has not yet been determined. We continue to work on this part.

4. Other Problems on Graphs and Digraphs

4.1. **Zero-Forcing.** [Collaboration with Sarah Behrens, Franklin Kenter, Thomas Mahoney, Keivan Monfared, Katy Nowak, and Michael Young started at the GRWC, [4]]

The zero-forcing model for the spread of infection on a graph was introduced by Burgarth and Giovannetti [6] and the AIM group [1]. With origins in the minimum rank problem on graphs, this model also has applications in quantum physics and networks.

For a fixed graph G, we start with a seed, which is a collection of initially infected vertices. Infection spreads dynamically according to the following rule: an infected vertex v spreads its disease to an uninfected neighbor u when u is the only uninfected neighbor of v. The zero-forcing number $Z(G)$ of a graph G is the minimum size of a seed that will spread infection to all of $V(G)$ given a sufficient amount of time. Davila and Kenter [8] conjectured $Z(G) \geq \delta + (\delta - 2)(g - 3)$, where δ is the minimum degree of G and g is the girth. We proved this bound for $g = 4$ and recently extended our methods to prove the conjecture for any girth. While it is NP-complete to compute the zero forcing number of a graph [10], we are interested in determining tight bounds for other classes of graphs.

4.2. **Graph Products.** [Collaboration with Richard Hammack, [11]]

We studied the algebraic concept of zero divisors in the context of digraphs. A digraph C is called a *zero divisor* if there exist non-isomorphic digraphs A and B for which $A \times C \cong B \times C$, where the operation is the direct product. In other words, C is a zero divisor if the cancellation property $A \times C \cong B \times C \Rightarrow A = B$ fails. Lovász [16] proved that C is a zero divisor if and only if it admits a homomorphism into a disjoint union of directed cycles of prime lengths. Thus any digraph C that is homomorphically equivalent to a directed cycle (or path) is a zero divisor. Given such a C and an arbitrary digraph A, we gave a method of computing all solutions X to the digraph equation $A \times C \cong X \times C$.

Our proof makes use of the digraph factorial which is defined as follows: For a digraph A, we let the vertices of $A!$ be the collection of permutations of $V(A)$. There is an edge from permutation α to permutation β in $A!$ provided for any $a, a' \in V(A)$, we have $aa' \in E(A)$ if and only if $\alpha(a)\beta(a') \in E(A)$. The factorial of a digraph, a relatively new construction introduced in [12], deserves more study since it led to such a nice solution in the zero divisor problem.

I particularly like this problem because of the algebraic flavor. In fact, the characterization that Lovász gave for zero divisors makes use of the Lovász Isotopy Lemma which is proven using category theory. I have taken a number of classes in universal algebra at USC and am an active participant in the Algebra and Logic seminar. With this background, I would like to explore more connections between algebra and combinatorics.

5. Dynamic Network Models

[Collaboration with Valentin Danchev, Beate Franke, and Pedro Reguiero started at the MRC]

Social networks often have natural communities which arise from geographic location, employment circles, interests, etc. Over time, these communities shift as life changes occur. We defined a dynamic model to track evolving community assignments with corresponding relationship changes that occur as nodes assimilate to their new communities.

Xu and Hero [24] developed a dynamic stochastic block model in which community membership changes over time. But at each time step, all edges are resampled. Snijder, van de Bunt, and Steglich [20] proposed a dynamic model where the relationships in each time step are influenced by the relationships in the previous time step. However, this model does not take into account the community structure. We combined the strengths of these two models to be able to track community membership while adding memory as we alter relationships.

Our current model has a fixed probability for a vertex to change communities and time-independent probabilities that edges will remain or be added from one time step to the next. We are interested in further refining our parameters to account for various properties of evolving community structure and using it to simulate and fit data.

As evidenced by the problems described above, my interests are centered in discrete mathematics, but I see value in gathering problems from within and outside mathematics that are relevant to a wider research community. Complexity, finite set systems and posets, extremal results of trees, graph games, graph products, and network models—I have some experience in 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 and learning from new colleagues. I look forward to the opportunity to supervise students on one or more of these projects. I am excited to continue my mathematical research and contribute to the life of a vibrant department.

References

[1] AIM Minimum Rank – Special Graphs Work Group. Zero forcing sets and the minimum rank of graphs. *Linear Algebra and its Applications*, **428**(7) (2008) 1628–1648.

[2] Zs. Baranyai. On the factorization of the complete uniform hypergraph, *Infinite and Finite Sets, Vol. 1.*, Proc. Coll. Keszthely, 1973, Amsterdam, Netherlands: North-Holland, (1975) 91–107.

[3] C.A. Barefoot, R.C. Entringer, L.A. Székely. Extremal values for ratios of distances in trees, *Discrete Applied Mathematics* **80** (1997) 37–56.
nullity of a graph. *Journal of Graph Theory*, **72**(2) (2013), 146–177.

[4] S. Behrens, F. Kenter, T. Mahoney, K. Monfared, K. Nowak, and M. Young. Lower bounds on the zero forcing number of graphs, *In preparation.*

[5] Beth, T., Algebraische Auflösungsalgorithmen für einige unendliche Familien von 3-Designs. *Le Matematiche* **29** (1974), 105–135.

[6] D. Burgarth and V. Giovannetti. Full control by locally induced relaxation. *Physical Review Letters*, **99**(10) (2007), 100501.

[7] P. Dankelmann, W. Goddard, C. Swart. The average eccentricity of a graph and its subgraphs, *Utilitas Mathematica* **65** (2004), 41–51.

[8] R. Davila and F. Kenter. Bounds for the zero-forcing number of graphs with large girth. *arXiv* 1406.0482v2 (2014).

[9] P.L. Erdős and L.A. Székely. On weighted multiways cuts in trees. *Mathematical Programming* **65**(1-3) (1994), 93–105.

[10] S. Fallat, K. Meagher, and B. Yang. On the complexity of the positive semidefinite zero forcing number. *arXiv* 1407:7017v1 (2014).

[11] R. Hammack and H. Smith. Zero divisors among digraphs. *Graphs and Combinatorics,* **30**(1) (2014), 171–181.

[12] R. Hammack and K. Toman, Cancellation of direct products of digraphs. *Discussiones Mathematicae Graph Theory*, **30** (2010), 575–590.

[13] C. Jordan, Sur les assemblages de lignes, *Journal für die reine und angewandte Mathematik* **70** (1869), 185–190.

[14] I. Miklós, Z.S. Kiss, and E. Tannier. Counting and sampling SCJ small parsimony solutions, *Theoretical Computer Science* **552** (2014), 83–98.

[15] I. Miklós and H. Smith. Complexity Results for the Single Cut-or-Join Model, *In preparation.*

[16] L. Lovász. On the cancellation law among finite relational structures. *Periodica Mathematica Hungarica* **1**(2) (1971), 145–156.

[17] H. Smith and L. A. Székely. Some remarks on Baranyai's Theorem. *Congressus Numerantium,* to appear 2014.

[18] H. Smith, L. A. Székely, H. Wang, and S. Yuan. Different middle parts of trees. *In preparation.*

[19] H. Smith, L. A. Székely, and H. Wang. Eccentricity in trees. *arXiv* 1408:5865, *Submitted to the Electronic Journal of Combinatorics* (2014+).

[20] T. A. B. Snijders, G. G. van de Bunt, and C. E. G. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks* **32**(1) (2010), 44–60.

[21] L. A. Székely, H. Wang. Extremal values of ratios: distance problems vs. subtree problems in trees, *Electronic Journal of Combinatorics* **20**(1) (2013), #P67, 1–20.

[22] H. Wang. The distances between internal vertices and leaves of a tree, *European Journal of Combinatorics*, 41 (2014), 79-99.

[23] H. Wang. The extremal values of the Wiener index of a tree with given degree sequence, *Discrete Applied Mathematics*, 156 (2008), 2647-2654.

[24] K. S. Xu and A. O. Hero. Dynamic stochastic blockmodels: statistical models for time-evolving networks. *Lecture Notes in Computer Science* **7812** (2013), 201–210.