UCLA
UCLA Electronic Theses and Dissertations

Title
Probability Models in Networks and Landscape Genetics

Permalink
https://escholarship.org/uc/item/99m7g2sv

Author
Ranola, John Michael Ordonez

Publication Date
2013

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Los Angeles

Probability Models in Networks and Landscape Genetics

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomathematics

by

John Michael Ordonez Ranola

2013

© Copyright by
John Michael Ordonez Ranola
2013

Abstract of the Dissertation

Probability Models in Networks and Landscape Genetics

by

John Michael Ordonez Ranola

Doctor of Philosophy in Biomathematics

University of California, Los Angeles, 2013

Professor Kenneth L. Lange, Chair

With the advent of massively parallel high-throughput sequencing, geneticists have the technology to answer many problems. What we lack are analytical tools. As the amount of data from these sequencers continues to overwhelm much of the current analytical tools, we must come up with more efficient methods for analysis. One potentially useful tool is the MM, majorize-minimize or minorize-maximize, algorithm.

The MM algorithm is an optimization method suitable for high-dimensional problems. It can avoid large matrix inversions, linearize problems, and separate parameters. Additionally it deals with constraints gracefully and can turn a non-differentiable problem into a smooth one. These benefits come at the cost of iteration.

In this thesis we apply the MM algorithm in the optimization of three problems. The first problem we tackle is an extension of random graph theory by Erdos. We extend the model by relaxing two of the three underlying assumptions, namely any number of edges can form between two nodes and edges form with a Poisson probability with mean dependent on the two nodes. This is aptly named a random multigraph.

The next problem extends random multigraphs to include clustering. As before, any number of edges can still form between two nodes. The difference is now the number of edges formed between two nodes is Poisson distributed with mean dependent on the two nodes along with their clusters.

For our last problem we place individuals onto the map using their genetic information. Using a binomial model with a nearest neighbor penalty, we estimate allele frequency surfaces for a region. With these allele frequency surfaces, we calculate the posterior probability that an individual comes from a location by a simple application of Bayes' rule and place him at his most probable location. Furthermore, with an additional model we estimate admixture coefficients of individuals across a pixellated landscape.

Each of these problems contain an underlying optimization problem which is solved using the MM algorithm. To demonstrate the utility of the models we applied them to various genetic datasets including POPRES, OMIM, gene expression, protein-protein interactions, and gene-gene interactions. Each example yielded interesting results in reasonable time.

The dissertation of John Michael Ordonez Ranola is approved.

Steve Horvath

Marc A. Suchard

Janet S. Sinsheimer

Kenneth L. Lange, Committee Chair

University of California, Los Angeles

2013

To my family ...

who have been a pillar of

support in all my endeavors.

Table of Contents

1 Introduction **1**

2 A Poisson Model for Random Multigraphs **4**

2.1 Motivation 4

2.2 Introduction 4

2.3 Background on the MM Algorithm 7

2.4 Methods 8

2.5 Results 10

2.5.1 *C. Elegans* Neural Network 10

2.5.2 Radiation Hybrid Gene Network 11

2.5.3 Protein Interactions via Literature Curation 13

2.5.4 Word Pairs and Letter Pairs 14

2.6 Conclusion 15

2.7 Tables and Figures 18

2.8 Appendix 18

2.8.1 Existence and Uniqueness of the Estimates 18

2.8.2 Convergence of the MM Algorithms 25

2.8.3 Log P-Value Approximations 26

2.8.4 Appendix Tables and Figures 26

3 Cluster and Propensity Based Approximation of a Network **37**

3.1 Abstract 37

3.1.1 Keywords 38

3.2 Background 38

3.2.1 Background: adjacency matrix and multigraphs 38
3.2.2 Background: correlation- and co-expression networks 39
3.3 Results and discussion 41
3.3.1 CPBA is a sparse approximation of a similarity measure 41
3.3.2 Objective functions for estimating CPBA 42
3.3.3 Example 1: Generalizing the random multigraph model 44
3.3.4 Example 2: Generalizing the conformity-based decomposition of a network 45
3.3.5 MM algorithm and R software implementation 46
3.3.6 Simulated clusters in the Euclidean plane 47
3.3.7 Simulated gene co-expression network 48
3.3.8 Real gene co-expression network application to brain data 48
3.3.9 OMIM disease and gene networks 49
3.3.10 Empirical comparison of edge statistics 54
3.3.11 Simulations for evaluating edge statistics 55
3.3.12 Hidden relationships between Fortune 500 companies 56
3.3.13 Relationship to other network models and future research 56
3.4 Conclusions 59
3.5 Methods 61
3.5.1 Maximizing the Poisson log-likelihood based objective function . . . 61
3.5.2 Minimizing the Frobenius norm based objective function 62
3.5.3 Model Initialization 63
3.5.4 Clustering algorithm 64
3.5.5 Quasi-Newton Acceleration 65
3.5.6 Estimating the number of clusters 66

3.6 Other . . . 67
3.6.1 Availability and requirements . . . 67
3.6.2 List of abbreviations . . . 67

4 Fast Spatial Ancestry via Flexible Allele Frequency Surfaces . . . 78
4.1 Abstract . . . 78
4.2 Introduction . . . 79
4.3 Results . . . 80
4.3.1 A Likelihood Ratio Criterion for SNP Selection . . . 80
4.3.2 Allele Frequency Surfaces . . . 82
4.3.3 Ancestral Origin Inference . . . 83
4.3.4 Estimating Proportions of Admixed Origins . . . 83
4.4 Methods . . . 89
4.4.1 A Likelihood Ratio Criterion for SNP Selection . . . 89
4.4.2 Allele Frequency Surface Estimation . . . 89
4.4.3 Localization of Unknowns . . . 91
4.4.4 Admixed Individuals . . . 92
4.5 Discussion . . . 93
4.6 Supplementary Results . . . 94

5 Future Work . . . 102
5.1 Landscape Genetics . . . 102
5.1.1 Spatial Haplotypes . . . 102
5.1.2 Landscape Weighting . . . 103
5.1.3 Individual vs. Group of Samples . . . 104
5.1.4 Sequence Data . . . 104

5.2 Landscape Measurements 105
5.2.1 Gaussian distribution 106
5.2.2 Poisson Model 107
5.2.3 Spatial-Temporal Measurements 108
5.3 Random Multigraphs and Barrier Identification 109
5.3.1 Bridge and Barrier Optimization 110

References **111**

List of Figures

2.1 Graph of a cluster of the radiation hybrid network significant connections ($p < 10^{-9}$) 18

2.2 Graph of a disjoint cluster of the HPRD dataset after analysis with our method using a cutoff of ($p < 10^{-6}$) 19

2.3 Graph of the significant connections ($p < 10^{-9}$) in the letter-pair network . . 20

2.4 Graph of *C. Elegans* neural network with a p-value of 10^{-6}. 27

2.5 Graph of the Radiation Hybrid network 36

3.1 Simulation providing a geometric interpretation of CPBA 69

3.2 Gene expression simulation results 70

3.3 Human brain expression data illustrate how CPBA can be interpreted as a generalization of WGCNA 71

3.4 OMIM disease network 72

3.5 OMIM Gene Network 73

3.6 OMIM CPBA versus PPP Analysis 76

3.7 Simulated CPBA versus PPP Analysis 77

4.1 Average distance between the geographic origin of the POPRES individuals and their SNPscape estimated origins as a function of the number of SNPs employed 80

4.2 Allele frequency surfaces generated by SNPscape with tuning parameter $\rho =$ 0.1 for the six most informative SNPs 81

4.3 Allele frequency surfaces generated by SPA for the six most informative SNPs 82

4.4 Average localization error for individuals based on leave-one-out cross validation using SNPscape ($\rho = 0.1$), SPA without SNP selection, and SPA with SNP selection. 84

4.5 Admixture coefficients for four simulated Europeans 87

4.6 Admixture coefficients for four simulated Europeans 88

4.7 A plot of the locations and sample sizes of the POPRES dataset. 95

4.8 Additional admixture coefficients for four simulated Europeans 96

4.9 Additional admixture coefficients for four simulated Europeans 97

4.10 Plot of the posterior probability of a three different individuals coming from each pixel using 50 SNPs with $\rho = 0.1$. 98

4.11 This figure shows the placement of all individuals back onto the map after using their data to generate allele frequency surfaces using various numbers of SNPs 99

4.12 This figure shows the placement of all individuals back onto the map after using their data to generate allele frequency surfaces using various numbers of SNPs continued 100

4.13 Additional estimated allele frequency surfaces for $\rho = 0.1$. 101

List of Tables

2.1 List of the 20 most significant connections of the *C. elegans* dataset. To the right of each pair appear the observed number of edges, the expected number of edges, and minus the log base 10 p-value. 28

2.2 Top 20 proteins with the most observed connections in the literature curated protein database. 29

2.3 The 20 proteins with the most significant connections ($p < 10^{-6}$) in the literature curated protein database. 30

2.4 BiNGO results of the small detached component around TP53 (Figure 2.2) in the literature curated protein database [MHK05] 31

2.5 Most significantly connected word pairs. 32

2.6 Words observed as a pair and never as singletons. 33

2.7 Most significantly connected letter pairs. 34

2.8 Convergence results for each of the 5 real datasets 35

3.1 Over-represented MeSH categories in the disease network. 51

3.2 Disease network top 15 significant connections CPBA. 52

3.3 Gene network top 20 significant connections CPBA. 53

3.4 Disease network top 15 significant connections PPP model. 74

3.5 Gene network top 20 significant connections PPP model. 75

3.6 Fortune 500 top 10 significant connections. 77

4.1 Comparison of localization by population 85

4.2 Accuracy of origin localization and run times for SNPscape, SCAT, and SPA for 100 SNPs 86

4.3 Comparison of SNPscape restricted to population pixels and Admixture . . . 86

Acknowledgments

I would never have been able to finish my dissertation without the guidance of my advisor and committee members, help from friends, and support from my family and wife. I would like to express my deepest gratitude to my advisor Ken Lange, for his patience, excellent guidance, care, and most of all his patience. Yes, I said patience twice. I know it took a lot of it to work with me at times, but he handled it well and was even able to teach me in the process. Thank you for sticking with me until the end. I would also like to thank my committee Janet Sinsheimer, Marc Suchard, and Steve Horvath for their guidance along the way.

I would like to thank Sangtae Ahn, Mary Sehl, and Desmond Smith for their work on Chapter 2 which is a version of the article [John Michael Ranola, Sangtae Ahn, Mary Sehl, Desmond Smith, and Ken Lange. "A Poisson model for random multigraphs." *Bioinformatics, 2010*]. Thank you Ken for your insight on the model and your guidance throughout this paper. Thank you Mary for finding and helping to analyze the literary data, and thank you to Desmond and Sangtae for helping with the analysis of the radiation hybrid network. Additionally, thanks to everyone in the group for the discussions we had on the analysis and writing up and reading the various parts of the paper.

I would also like to thank Steve Horvath and Peter Langfelder for their work on Chapter 3 which is a version of the article [John Michael Ranola, Peter Langfelder, Kenneth Lange, and Steve Horvath. "Cluster and propensity based approximation of a network." *BMC systems biology, 2013*]. Once again I want to thank Ken for his insight and guidance throughout this problem. Thank you Peter for your help in analyzing the human brain expression data and in creating the PropClust package for R. Thank you Steve for your guidance in developing the model and selecting appropriate data sets. Additionally, thanks to everyone in the group for writing and reading the paper.

For Chapter 4, I would like to thank John Novembre for his help in developing the model and finding appropriate data. Hopefully we can publish this soon.

I would like to thank my parents Rene and Cynthia, my brother Ryan, and my sister

Jaimee. They were always supporting me and encouraging me in their own way. Finally, I would like to thank my wife Antoinette. Graduate school was a rollercoaster full of ups, downs, and loops at times, but with her there it was bearable. I couldn't have finished the ride without her.

Thank you all so much.

Vita

2004	REU: Mathematical Biology, Penn State University, Erie PA
2005	Murdock Internship in Biomechanics, University of Portland, Portland OR
2005	REU: Mathematics of Flight, Kansas State University, Manhattan KS
2006	B.S. (Mathematics and Biology), University of Portland, Portland OR
2008	M.S. (Biomathematics), University of California Los Angeles, Los Angeles CA
2009–present	Research Assistant, Biomathematics Department, University of California Los Angeles, Los Angeles CA

Publications and Presentations

Ranola J, Novembre J, and Lange K. *Genographical Estimation and Projection.* In progress.

Ranola J, Langfelder P, Lange K, and Horvath S. *Cluster and propensity based approximation of a network.* BMC Systems Biology 2013; 7:21.

Ranola JM, Ahn S, Sehl M, Smith DJ, and Lange K. *A Poisson model for random multigraphs.* BMC Bioinformatics 2010; 26(16):2004-11.

Ranola J, Tobalske B, Warrick D, and Powers D. *Circulation in the wake of the flying hummingbird: Effects of thresholding and vortex decay.* Int. Comp. Biol., 45, 1181.

WNAR/IMS Student Speaker, UCLA, Summer 2013

Biomathematics 210: Optimization methods in Biology Guest Speaker, UCLA, Fall 2009

Systems & Integrative Biology Retreat Speaker, UCLA, Winter 2009

Student Speaker for the Mathematical Association of America Northwest Regional Meeting, University of Puget Sound, Spring 2005

Featured Student Speaker for the 14th Regional Conference on Undergraduate Research of the Murdock College Science Research Program, Northwest Nazarene University, Fall 2005

CHAPTER 1

Introduction

During the past decade we have seen great hurdles surmounted in the field of genomics. The first draft of the human genome [LLB01], the development of a haplotype map of the human genome [GBH03], and the official completion of the human genome project's goal of a completed human genome [CLR04], heralded the age of genomics [Wal01]. Of course, as with every age, each surmounted hurdle only reveals more hurdles to conquer. With the recent advent of massively parallel high throughput sequencers [QSC12, TMF09] we now have the physical tools needed for tackling harder problems. Unfortunately, we still lack the analytical and computational tools. The new sequencers has brought a plethora of data that is orders of magnitude larger than what we were used to; in many cases it is far above what current computers and analysis methods can handle. In order to understand the data properly and advance to the next milestone, new analytical tools need to be developed which can handle the large amounts of data in reasonable time. One useful tool for doing so is the MM, majorize-minimize or minorize-maximize, algorithm [OR00, DH77, LHY00, HL00].

The MM algorithm is a method for optimization. The beauty of the MM algorithm is that it substitutes a simple optimization problem for a difficult one. The idea behind the MM algorithm is to substitute the current function, $f(\theta)$ with a surrogate which majorizes or minorizes it. The new function must be tangent to the function at the current iterate θ^n, and it must dominate it elsewhere. In symbols

$$g(\theta^n|\theta^n) = f(\theta^n)\,, \text{ and}$$
$$g(\theta|\theta^n) \geq f(\theta) \text{ for all } \theta.$$

The next iterate, θ^{n+1}, is then chosen to minimize the surrogate function $g(\theta|\theta^n)$ rather than the original. This process is iterated until convergence. One benefit of the MM algorithm

is its numerical stability, indeed the descent property guarantees that iterations are always getting better. This can be shown through the inequality

$$\begin{aligned} f(\theta^{n+1}) &\leq g(\theta^{n+1}|\theta^n) \\ &\leq g(\theta^n|\theta^n) \\ &= f(\theta^n)\,, \end{aligned}$$

where the first inequality holds due to the dominance requirement, the second holds due to θ^{n+1} minimizing the surrogate function $g(\theta|\theta^n)$, and the last holds due to the tangency requirement.

Like all tools the MM algorithm has drawbacks. One is that, like Newton's method, it is unable to distinguish between local and global minima. The second is that their convergence rate is often slow in the neighborhood of the minimum point. Though there is no real good solution for the first drawback other than multiple starting points, the second drawback can be alleviated by schemes such as quasi-Newton acceleration [ZAL11]. In the coming chapters we demonstrate the utility of the MM algorithm and quasi-Newton acceleration through three problems.

The first model we tackle is an extension of the well researched random graph theory [ER59, BA99]. It has been shown that the simple model is often too rigid to capture real-world networks [AB02, Str01]. To alleviate this, we present a random multigraph model [RAS10]. In it, we relax two of the three original assumptions and llow any number of edges to form between nodes. We also assume edges form with a Poisson probability with a mean $p_i p_j$, where p_i is the propensity associated with node i. The third requirement of independent edge formation is kept intact. These assumptions give rise to a probability model which we are easily able to maximize using the MM algorithm and accelerate with quasi-Newton acceleration. Additionally, we present a directed multigraph approach which gives each node an incoming propensity p_i and an outgoing propensity q_i with the mean number of edges from i to j now being $q_i p_j$. This is also optimized via the MM algorithm and accelerated. To show the value of the model, we apply it to a neural network, a gene network, a literature curated protein interaction network, and a literary network based on

some of Shakespeare's plays.

We extend the multigraph model even further by including clustering. Additionally, we add a least squares form to include weighted networks in analysis [RLL13]. In clustering model, each node belongs to a cluster, and each cluster has some propensity to interact with other clusters. The mean number of edges between nodes i and j is now $A_{c_i c_j} p_i p_j$, where p_i is the propensity of node i, c_i is the cluster of node i, and $A_{c_i c_j}$ is the intercluster adjacency between the cluster of i and j. This again leads to a likelihood or a least squares criterion that is optimized via the MM algorithm and accelerated. We apply this method to gene expression data [OKI08], a bipartite network of diseases and genes from the Online Mendelian Inheritance of Man (OMIM)[HSA05], and a network created from shared board members of Fortune 500 companies.

For our final problem, we placed individuals onto the geographic map using their genetic information. Although this problem has been tackled before in various ways [NJB08, YNE12, WSC04], there was room for improvement. In our solution to the problem, we began by pixellating the region of interest. We then used a binomial model with a nearest neighbor penalty to estimate the allele frequencies at all pixels. The allele frequencies of pixels with data are mainly driven by the binomial model while the penalty allowed us to estimate the allele frequencies of those pixels without data by borrowing strength from their neighbors. The penalized loglikelihood was optimized via the MM algorithm and accelerated. We applied the model to the POPRES dataset, consisting of 1387 individuals from 37 countries mapped at nearly 200,000 SNPs [NBK08], and compared it to the other methods. Furthermore, utilizing the estimated allele frequency surfaces, we presented a model to place admixed individuals onto the map. This model was also optimized via the MM algorithm and accelerated. We applied the admixed model to admixed individuals simulated from the POPRES dataset. Our results in both the unmixed and the admixed cases are encouraging.

CHAPTER 2

A Poisson Model for Random Multigraphs

2.1 Motivation

Biological networks are often modeled by random graphs. A better modeling vehicle is a multigraph where each pair of nodes is connected by a Poisson number of edges. In the current model the mean number of edges equals the product of two propensities, one for each node. In this context it is possible to construct a simple and effective algorithm for rapid maximum likelihood estimation of all propensities. Given estimated propensities, it is then possible to test statistically for functionally connected nodes that show an excess of observed edges over expected edges. The model extends readily to directed multigraphs. Here propensities are replaced by outgoing and incoming propensities.

2.2 Introduction

Random graph theory has proved vital in modeling the internet and constructing biological and social networks. In the original formulation of the theory by Erdös and Rényi, there are three key assumptions: (a) a graph exhibits at most one edge between any two nodes, (b) the formation of a given edge is independent of the formation of other edges, and (c) all edges form with the same probability [ER59, ER60]. There is general agreement that this simple model is too rigid to capture many real-world networks [AB02, Str01]. The surveys [BA99, Dur07, NSW01] summarize some of the elaborations and applications of two generations of scholars, with emphasis on power laws, phase transitions, and scale-free networks. In the current paper we study a multigraph extension of the Erdös-Rényi model

appropriate for very large networks. Our model specifically relaxes assumptions (a) and (c). With appropriate alternative assumptions in place, we derive and illustrate a novel maximum likelihood algorithm for estimation of the model parameters. With these parameters in hand, we are then able to find statistically significant connections between pairs of nodes.

In practice many graphs are derived from multigraphs. To simplify analysis, the multiple edges between two nodes of a multigraph are collapsed to a single edge. The movie star example in reference [NSW01] is typical. In the movie star graph, two actors are connected by an edge when they appear in the same movie. Some actor pairs will appear in a movie mostly by chance. Other actor pairs will be connected by multiple edges because they are intrinsically linked. Classic pairs such as Abbot and Costello, Loy and Powell, and Lewis and Martin come to mind.

The well-studied neural network of *C. elegans* is a prime biological example. Here neuron pairs are connected by multiple synapses. Because collapsing edges wastes information, it is better to tackle the multiplicity issue directly. Thus, we will deal with random multigraphs. For our purposes, these exclude loops and fractional edge weights. Instead of a Bernoulli number of edges between any two nodes as in the Erdös and Rényi model, we postulate a Poisson number of edges. This choice can be viewed as unnecessarily restrictive, but it is worth recalling that a Poisson distribution can approximate a binomial or normal distribution. Furthermore, the Poisson assumption allows an arbitrary mean number of edges.

In relaxing assumption (c) above, we want to introduce as few parameters as possible but still capture the capacity of some nodes to serve as hubs. Thus, we assign to each node i a propensity p_i to form edges. The random number of edges X_{ij} between nodes i and j is then taken to be Poisson distributed with mean $p_i p_j$. Node pairs with high propensities will have many edges, pairs with low propensities will have few edges, and pairs with one high and one low propensity will have intermediate numbers of edges. Later we will show that these choices promote simple and rapid estimation of the propensities. Another virtue of the model is that it generalizes to directed graphs where arcs replace edges. For directed graphs, we postulate an outgoing propensity p_i and an incoming propensity q_i for each node i. The number of arcs X_{ij} from i to j is taken to be Poisson distributed with mean $p_i q_j$. In

the directed version of the model, the two random variables X_{ij} and X_{ji} are distinguished. In accord with assumption (b), the random counts X_{ij} in either model are taken to be independent.

Protein and gene networks can involve tens of thousands of nodes. Estimation of propensities under the Poisson multigraph model for such networks is consequently problematic. Standard algorithms for parameter estimation such as least squares, Newton's method, and Fisher scoring require computing, storing, and inverting large Hessian matrices. Such actions are not really options in high-dimensional problems. One of the biggest challenges in the present paper is crafting an alternative estimation algorithm that remains viable in high dimensions. Fortunately, the MM (minorize-maximize) principle [Lan04, LHY00] allows one to design a simple iterative algorithm for the random multigraph model. Large matrices are avoided, and convergence is reasonably fast. In the appendix we prove that the new MM algorithm converges to the global maximum of the likelihood.

Another strength of the model is that it permits assessment of statistical significance. In other words, it helps distinguish random connectivity from functional connectivity. The basic idea is very simple. Every edge count X_{ij} is Poisson distributed with a parameterized mean. If we substitute estimated propensities for theoretical propensities, then we can estimate the mean and therefore approximate the tail probability $p = \Pr(X_{ij} \geq x_{ij})$ associated with the observed number of edges x_{ij} between two nodes i and j. The smaller this probability, the less likely these edges occur entirely by chance. For instance in the movie star example, the actor pair Abbot and Costello would be flagged as significant in any representative data set of their era. In less obvious examples, discerning functionally connected pairs is more challenging. In the appendix we show how to approximate very low p-values under the Poisson distribution.

To test the model, we analyze five real data sets. Three of these are biological and involve undirected graphs. The first is the neural network of *C. elegans* [WS98, WST86] already mentioned. The second is a network obtained by subjecting a panel of radiation hybrids to gene expression measurements [AWP09, PAB08]. In the network two genes are connected by an edge if a marker significantly regulates the expression levels of both genes

in the clones of the panel. Our third biological example involves interacting proteins taken from the curated Human Protein Reference Database [PGK09]. For directed graphs we turn to literary analysis of a subset of Shakespeare's plays. Here we look at letter pairs and word pairs. Every time the first letter of a pair precedes the second letter of a pair in a word, we introduce an arc between them. Likewise, every time the first word of a pair precedes the second word of a pair in a sentence, we introduce an arc between them. Other applications such as monitoring internet traffic come immediately to mind but will not be treated here.

Let us stress the exploratory nature of the Poisson multigraph model. Its purpose is to probe large data sets for hidden structure. Identifying hub nodes and node pairs with excess edges are primary goals. The fact that the model is at best a cartoon does not eliminate these possibilities. For example, even if we do not take the p-values generated by the model seriously, they can still serve to rank important node pairs for further investigation and experimentation. Computational biology is full of compromises between realistic models and computational feasibility.

Before tackling these specific examples, we will briefly review the MM principle and lay out the details of the model. Once this foundation is in place, we show how a simple inequality drives the optimization process. The MM principle is designed to steadily increase the loglikelihood of the model given the data. This ascent property is the key to understanding how the algorithm operates.

2.3 Background on the MM Algorithm

As we have already emphasized, the MM algorithm is a principle for creating algorithms rather than a single algorithm. There are two versions of the MM principle, one for iterative minimization and another for iterative maximization. Here we deal only with the maximization version. Let $L(\mathbf{p})$ be the objective function we seek to maximize. An MM algorithm involves minorizing $L(\mathbf{p})$ by a surrogate function $g(\mathbf{p} \mid \mathbf{p}^n)$ anchored at the current iterate

$\mathbf{p}^n$ of a search. Minorization is defined by the two properties

$$L(\mathbf{p}^n) = g(\mathbf{p}^n \mid \mathbf{p}^n) \tag{2.1}$$

$$L(\mathbf{p}) \geq g(\mathbf{p} \mid \mathbf{p}^n), \qquad \mathbf{p} \neq \mathbf{p}^n. \tag{2.2}$$

In other words, the surface $\mathbf{p} \mapsto g(\mathbf{p} \mid \mathbf{p}^n)$ lies below the surface $\mathbf{p} \mapsto L(\mathbf{p})$ and is tangent to it at the point $\mathbf{p} = \mathbf{p}^n$. Construction of the surrogate function $g(\mathbf{p} \mid \mathbf{p}^n)$ constitutes the first M of the MM algorithm. In the second M of the algorithm, we maximize the surrogate function $g(\mathbf{p} \mid \mathbf{p}^n)$ rather than $L(\mathbf{p})$. If $\mathbf{p}^{n+1}$ denotes the maximum point of $g(\mathbf{p} \mid \mathbf{p}^n)$, then this action forces the ascent property $L(\mathbf{p}^{n+1}) \geq L(\mathbf{p}^n)$. The straightforward proof

$$L(\mathbf{p}^{n+1}) \geq g(\mathbf{p}^{n+1} \mid \mathbf{p}^n) \geq g(\mathbf{p}^n \mid \mathbf{p}^n) = L(\mathbf{p}^n),$$

reflects definitions (2.1) and (2.2) and the choice of $\mathbf{p}^{n+1}$. The ascent property is the source of the MM algorithm's numerical stability. Strictly speaking, it depends only on increasing $g(\mathbf{p} \mid \mathbf{p}^n)$, not on maximizing $g(\mathbf{p} \mid \mathbf{p}^n)$.

The celebrated EM algorithm [DLR77] is a special case of the MM algorithm [LHY00, Lan04]. The EM algorithm always relies on some notion of missing data. Discerning the missing data in a statistical problem is sometimes easy and sometimes hard. In our Poisson graph model, it is unclear what constitutes the missing data. In contrast, derivation of a reliable MM algorithm is straightforward but ad hoc. Readers wanting a more systematic derivation are apt to be disappointed. In our defense it is possible to codify several successful strategies for constructing surrogate functions [LHY00, HL04, Lan04].

2.4 Methods

Consider a random multigraph with m nodes labeled $1, 2, \ldots, m$. A random number of edges X_{ij} connects every pair of nodes $\{i, j\}$. We assume that the X_{ij} are independent Poisson random variables with means μ_{ij}. As a plausible model for ranking nodes, we take $\mu_{ij} = p_i p_j$, where p_i and p_j are nonnegative propensities. The loglikelihood of the observed edge counts

$x_{ij} = x_{ji}$ amounts to

$$\begin{aligned} L(\mathbf{p}) &= \sum_{\{i,j\}} (x_{ij} \ln \mu_{ij} - \mu_{ij} - \ln x_{ij}!) \\ &= \sum_{\{i,j\}} [x_{ij}(\ln p_i + \ln p_j) - p_i p_j - \ln x_{ij}!]. \end{aligned}$$

Inspection of $L(\mathbf{p})$ shows that the parameters are separated except for the products $p_i p_j$. To achieve full separation of parameters in maximum likelihood estimation, we employ the majorization

$$p_i p_j \leq \frac{p_j^n}{2p_i^n} p_i^2 + \frac{p_i^n}{2p_j^n} p_j^2$$

with the superscript n indicating iteration. Observe that equality prevails when $\mathbf{p} = \mathbf{p}^n$. This majorization leads to the minorization

$$\begin{aligned} L(\mathbf{p}) &\geq \sum_{\{i,j\}} [x_{ij}(\ln p_i + \ln p_j) - \frac{p_j^n}{2p_i^n} p_i^2 - \frac{p_i^n}{2p_j^n} p_j^2 - \ln x_{ij}!] \\ &= g(\mathbf{p} \mid \mathbf{p}^n). \end{aligned}$$

Maximization of $g(\mathbf{p} \mid \mathbf{p}^n)$ can be accomplished by setting

$$\frac{\partial}{\partial p_i} g(\mathbf{p} \mid \mathbf{p}^n) = \sum_{j \neq i} \frac{x_{ij}}{p_i} - \sum_{j \neq i} \frac{p_j^n}{p_i^n} p_i. = 0$$

The solution

$$p_i^{n+1} = \sqrt{\frac{p_i^n \sum_{j \neq i} x_{ij}}{\sum_{j \neq i} p_j^n}} \tag{2.3}$$

is straightforward to implement and maps positive parameters to positive parameters. When edges are sparse, the range of summation in $\sum_{j \neq i} x_{ij}$ can be limited to those nodes j with $x_{ij} > 0$. Observe that these sums need only be computed once. The partial sums $\sum_{j \neq i} p_j^n = \sum_j p_j^n - p_i^n$ require updating the full sum $\sum_j p_j^n$ once per iteration. A similar MM algorithm can be derived for a Poisson model of arc formation in a directed multigraph. We now postulate a donor propensity p_i and a recipient propensity q_j for arcs extending from node i to node j. If the number of such arcs X_{ij} is Poisson distributed with mean $p_i q_j$, then under independence we have the loglikelihood

$$L(\mathbf{p}, \mathbf{q}) = \sum_i \sum_{j \neq i} [x_{ij}(\ln p_i + \ln q_j) - p_i q_j - \ln x_{ij}!]$$

With directed arcs the observed numbers x_{ij} and x_{ji} may differ. The minorization

$$\begin{aligned} L(\mathbf{p},\mathbf{q}) &\geq \sum_i \sum_{j\neq i} [x_{ij}(\ln p_i + \ln q_j) \\ &\quad -\frac{q_j^n}{2p_i^n}p_i^2 - \frac{p_i^n}{2q_j^n}q_j^2 - \ln x_{ij}!] \end{aligned}$$

now yields the MM updates

$$p_i^{n+1} = \sqrt{\frac{p_i^n \sum_{j\neq i} x_{ij}}{\sum_{j\neq i} q_j^n}}, \quad q_j^{n+1} = \sqrt{\frac{q_j^n \sum_{j\neq i} x_{ji}}{\sum_{j\neq i} p_i^n}}.$$

Again these are computationally simple to implement and map positive parameters to positive parameters. It is important to observe that the loglikelihood $L(\mathbf{p},\mathbf{q})$ is invariant under the rescaling cp_i and $c^{-1}q_j$ for a positive constant c and all i and j. This fact suggests that we fix one propensity and omit its update. To derive a reasonable starting value in the undirected multigraph model, we maximize $L(\mathbf{p})$ under the assumption that all p_i coincide. This gives the initial values

$$p_k^0 = \sqrt{\frac{\sum_{\{i,j\}} x_{ij}}{\binom{m}{2}}}.$$

The same conclusion can be reached by equating theoretical and sample means. In the directed multigraph model, we maximize $L(\mathbf{p},\mathbf{q})$ subject to the restriction that all p_i and q_j coincide. Now we have

$$p_k^0 = q_k^0 = \sqrt{\frac{\sum_i \sum_{j\neq i} x_{ij}}{m(m-1)}}.$$

Note that the fixed parameter is determined by this initialization.

2.5 Results

2.5.1 *C. Elegans* Neural Network

The neural network of *C. elegans* is a classic dataset first studied by [WST86] and later by [WS98]. In their paper, White et al. were able to obtain high resolution electron microscopic images. This allowed them to identify all the synapses, map all the connections, and to work

out the entire neuronal network of the worm. To use all known connections in our analysis, we add as edges the electric junctions and neuromuscular junctions observed by Chen et al [CHC06]. For consistency we disregard the directionality of the chemical synapses. In our opinion, the flexibility of the model in accepting different definitions of edges should be viewed as a strength. We declare a connection between two neurons i and j to be functionally significant when $\Pr(X_{ij} \geq x_{ij}) \leq 10^{-6}$. Figure 1 in the Appendix depicts the network.

As recorded in Table 2.1, many of the most significant connections extend between motor neurons. The model also captures the bilateral symmetry between the right and left sides of the worm. Thus, the connections between the pairs RIPR-IL2VR and RIPL-IL2VL and between OLLL-AVEL and OLLR-AVER are all significant. Note that an L or an R at the end of a neuron's name signifies the left and right side, respectively. The right neuron PDER appears twice on the top 50 list and and its left counterpart PDEL is missing, but both have the same number of significant edges overall. Although these dual connections are highlighted as about equally significant in our analysis, the corresponding propensity estimates show a left-right imbalance. The cause of these slight departures from bilateral symmetry is obscure. In any event, the model is subtle enough to distinguish between high edge counts and significant edge counts. Thus, even though one pair of nodes may have more edges than another pair, it does not necessarily imply that the first pair is more significantly connected than the second pair.

2.5.2 Radiation Hybrid Gene Network

Radiation hybrids were originally devised as a tool for gene mapping [GH75] at the chromosome level. The detailed physical maps they ultimately provided [KM90] served as a scaffolding for sequencing the entire human genome. To construct radiation hybrids, one irradiates cells from a donor species. This fragments the chromosomes and kills the vast majority of cells. A few donor cells are rescued by fusing them with cells of a recipient species. Some of the fragments, say 10%, get translocated or inserted into the chromosomes of the recipient species. The hybrid cells have no particular growth advantage over the more

numerous unfused recipient cells. However, if cells from the recipient cell line lack an enzyme such as hypoxanthine phosphoribosyl transferase (HPRT) or thymidine kinase (TK), both the unfused and the hybrid cells can be grown in a selective medium that eliminates the unfused recipient cells. This selection process leaves a few hybrid cells, and each of the hybrid cells serves as a progenitor of a clone of identical cells. Each clone contains a random subset of the genome of the donor species. The presence or absence of a particular short region can be assayed by testing for a donor marker in that region. A given donor marker is present in a given clone in 0, 1, or 2 copies.

It turns out that one can exploit radiation hybrids to map QTLs (quantitative trait loci). We measured the log intensities of 232,626 aCGH (array comparative genomic hybridization) markers and 20,145 gene expression levels in each of 99 mouse-hamster radiation hybrids [AWP09, PAB08]. In this case a mouse served as the donor and a hamster as the recipient. We then regressed the mouse gene expression levels on the mouse copy numbers recorded for each of the mouse markers. Altogether this amounts to about 5×10^9 separate linear regressions. We constructed a multigraph from the data by analogy with the movie star example, with genes corresponding to actors and markers to movies. An edge is added between two genes if both genes showed statistically significant dependence on the marker at the level $p \leq 10^{-9}$. This strict p-value cutoff was chosen to produce an easily visualized graph. Because the aCGH markers densely cover the mouse genome, a quasi-peak finding algorithm was used to delete the excess edges occurring under a common linkage peak. Figure 2 in the Appendix depicts the full network. Here node size is proportional to estimated propensity, and edge darkness is proportional to significance. Red edges are the most significant. Even with a very stringent significance level and elimination of edges by peak finding, there are still 729,169 significant connections.

Figure 2.1 shows an interesting subnetwork with highly significant edges, genes (nodes) of large propensity, and genes with related functions. The Dishevelled 1 (Dvl1) member of this subnetwork is part of the wingless/Int (Wnt) signaling pathway. The Wnt pathway has a reciprocal signaling relationship with the hedgehog pathway, which requires oxysterols for optimal function [CS06]. The Wnt hedgehog connection is important in stem cell renewal.

Interestingly, oxysterol binding protein-like 3 (Osbpl3) is a member of the subnetwork as well as Dvl1. Furthermore, the subnetwork contains two membrane associated proteins: mucolipin 3(Mcoln3), a cation channel protein [CS08], and aquaporin 2(Aqp2), a water channel protein [CA09]. An emerging theme in cancer research is the notion of evolving genetic networks [MMS08]. Networks constructed using the Poisson multigraph model can robustly identify unexpected connections with known oncogene pathways such as the Wnt pathway. These connections may ultimately suggest novel therapeutic strategies.

2.5.3 Protein Interactions via Literature Curation

With the advent of high throughput experimentation, an enormous mass of information on protein interactions has accumulated. Because there was initially no universal format for presenting interactions, many of the early discoveries were useful only to the originating labs. This bottleneck forced coordination and eventually the construction of unified databases with fixed formats combining all of the published information. A notable example of this process of curation is the Human Protein Reference Database [PGK09]. We downloaded Release 7 of the database and analyzed it with the random multigraph model.

Several interesting features of the data emerge under a p-value cutoff of 10^{-6}. For instance, the protein with the most observed edges, TP53, turns out to be different from the protein with the most significant edges, Stat3. In fact, none of the top five proteins ranked by the most observed edges are in the top five proteins ranked by the most significant edge counts. Thus, the hub nodes of the raw data differ sharply from the hub nodes of the processed data. The two most extreme cases, YWHAG and CREBBP, have no significant edge counts despite being ranked fourth and fifth based on observed edges (See Tables 2.2 and 2.3). One should be cautious in interpreting such results because molecular experiments are hypothesis driven and generate very biased data. The value of looking for significance is that it turns up hidden structure, not that it calls into question known structure.

When we cluster proteins by significant edge counts, the TP53 protein is especially interesting. Consider the small component containing TP53 shown in Figure 2.2. We analyzed

this cluster using the BiNGO addition to Cytoscape [MHK05]. BiNGO computes the probability that x or more genes in a given set of genes shares the same GO (gene ontology) category. Altogether we found 30 significant GO categories with $p < 10^{-6}$; most of these categories are listed in Table 2.4. These results dramatically illustrate the role of TP53 in regulating the cell cycle by (a) activating DNA repair proteins, (b) arresting the cell cycle at the G_1/ S checkpoint to permit repair, and (c) initiating apoptosis in extreme circumstances.

2.5.4 Word Pairs and Letter Pairs

Identifying frequently used word pairs in literary texts can be useful in problems of literary attribution and in the identification of word fossils. Vocabulary richness and frequencies of sets of words have been studied in many different literary contexts using a variety of methods, including, for example, Bayesian analysis and machine learning to determine authorship of the Federalist papers [MW84, HF95], and likelihood ratio tests to study the *Pearl* poems [MW83]. Recent investigations of long texts [BRM09] have called into question Zipf's law [Zip32], which postulates that the frequency of any word is inversely proportional to its rank in usage. Here we apply the Poisson model of graph connectivity to study pairs of words used consecutively in a set of Shakespeare's plays.

Our version of word pair analysis begins by scanning a literary work and creating a dictionary of words found in the text. An arc is drawn between two consecutive words, from the first word to the second word of the text, provided the words are not separated by a punctuation mark. The number of arcs between an ordered pair of words is counted and stored in a square matrix with dimensions equal to the number of unique words in the text. We chose seven of Shakespeare's plays, *All's Well that Ends Well*, *As You Like It*, *Julius Caesar*, *King Lear*, *Macbeth*, *Measure for Measure*, and *Titus Andronicus*, concatenated them, and analyzed them as a whole. Contractions such as "o'er" and "ta'en" were replaced by the corresponding full words, "over" and "taken", respectively. We retained in our analysis word-pairs constituting character names.

We calculated the observed frequency of each word pair. Based on the directed random

multigraph model described in the Methods section, we estimated the outgoing and incoming propensities for each word along with expected frequencies and p-values for each word pair. Table 2.5 lists the most connected word pairs in the text ranked by their p-values. This set is dominated by phrases that are commonly used in the language of the day, such as "I am" and "my lord", and by character names, such as "Lady Macbeth" and "Second Lord", in each play.

One can identify several word pairs whose members almost never occur separately by examining the ratio $x_{ij}/(\hat{p}_i\hat{q}_j)$ of observed to expected word pair frequencies. Table 2.6 lists several examples ranked by this index. These word-pair fossils are dominated by a few phrases still in common use such as "pell mell" and "tick tack" as well as various Latin and Italian phrases, such as "et tu Brute", and other strange phrases specific to the context of particular plays, such as "boarish fangs" and "rustic revelry."

In addition, we studied pairs of letters encountered consecutively in the combined text of the Shakespearean plays. Figure 2.3 depicts the letter-pair connections using a very stringent p-value of 10^{-19} for display purposes. Table 2.7 lists the same results in tabular form. The two most significant pairs are "th" and "he". One would expect much more stability over time of letter-pair usage than word-pair usage. This contention is borne out by our separate analysis of the novel *Jane Eyre* by Charlotte Bronte.

2.6 Conclusion

Multigraphs are inherently more informative than ordinary graphs, and random multigraphs offer rich possibilities for modeling biological, social, and communication networks. Our applications are meant to be illustrative rather than exhaustive. Graphical models will surely grow in importance as research laboratories and corporations gather ever larger data sets and hire ever more computer scientists and statisticians to mine them. The Poisson model has many advantages. It is flexible enough to capture hub nodes and functional connectivity, generalizes to directed graphs, and sustains an MM estimation algorithm capable of handling enormous numbers of nodes.. It is also very quick computationally as measured by total

iterations and total time until convergence. A glance at Table 1 of the Appendix suggests that 20 to 30 iterations suffice for convergence. To thrive, data mining must balance model realism with model computability. In our opinion, the Poisson model achieves this end. Of course, other distributions for edge counts could be tried, for instance the binomial or the negative binomial, but they would be even less well motivated and less adapted to fast estimation.

It is natural to place our advances in the larger context of applied random graph theory. For instance, early on social scientists married latent variable models and random networks [HL81]. Stochastic blockmodels assign nodes either deterministically or stochastically to latent classes [ABF08, HLL83, NL07, NS01, WW87]. Alternatively, a latent distance model sets up a social space and estimates the distances between node pairs in this space [HRH02]. It is possible to combine features of both latent class and latent distance models in a single eigenmodel [Hof07]. The "attract and introduce" model is another helpful elaboration [FDC09]. None of these models focuses on multigraphs. Furthermore, most classical applications involve networks of modest size. However, under the stimulus of large internet data sets, the field of random networks is in rapid flux. Going forward it will be a challenge to turn the rising flood of data into useful information. Importing more of the social science contributions into biological research may pay substantial dividends.

In practice, most large networks contain an excess of weak interactions. The radiation hybrid data is typical in this regard. To sift through the data, it is helpful to focus on hub nodes and strong interactions. The Poisson multigraph model provides a rigorous way of doing so. The model's flexibility in allowing different sorts of edges is appealing if not taken to extremes. When confidence in edge assignment varies widely across edge definitions, a weighted graph model might be a better modeling device than a multigraph model. However, converting a multigraph to a weighted graph has its own problems. For instance, there is more than one way to make the conversion. An even bigger disadvantage of weighted graph models is their tendency to ignore the stochastic nature of node formation. This is a hindrance in assessing functional connections and suggests an opportunity for more nuanced modeling. To be competitive with Poisson multigraphs, a good stochastic model

for weighted graphs should support fast estimation of parameters. One substitute for Poisson randomness is to condition on the degree of each node [CL02]. Within these constraints, one can randomize edge placement. This perspective lends itself to permutation testing but not to parameter estimation [MS02]. Unfortunately, the computational cost of generating the required permutations limits the chances for approximating very small p-values and hence ranking connections by p-values.

The random multigraph model raises as many questions as it answers. How closely is it tied to the Poisson distribution? How closely is it tied to the propensity parameterization of edge means? Can predictors be incorporated that determine propensities? More importantly, what applications would benefit from this sort of modeling? We are content to raise these issues, with the hope that other computational and mathematical scientists can be enlisted over time to resolve them and related problems beyond our current understanding.

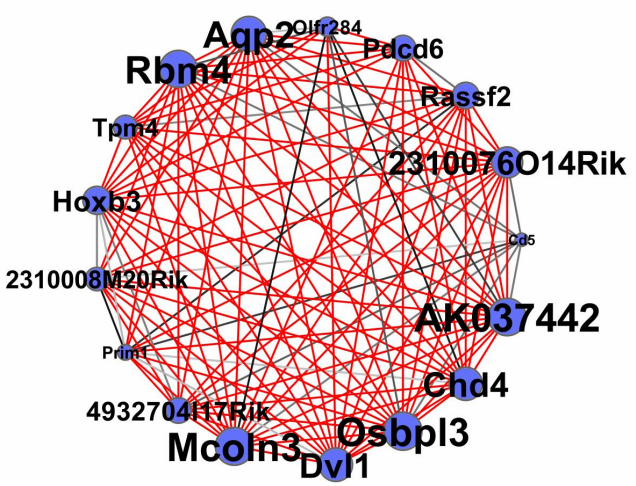


Figure 2.1: Graph of a cluster of the radiation hybrid network with significant connections ($p < 10^{-9}$). In this graph, node size is proportional to a node's estimated propensity. Also, the darker the edge, the more significant the connection; red lines highlight the most significant connections. Edges between this cluster and the rest of the network were removed for clarity.

2.7 Tables and Figures

2.8 Appendix

2.8.1 Existence and Uniqueness of the Estimates

The body of the paper takes for granted the existence and uniqueness of the maximum likelihood estimates. These more subtle questions can be tackled by reparameterizing. Before we do so, let us dismiss the exceptional cases where a node has no edges. If this condition holds for node i, then in the undirected graph model the value $p_i = 0$ maximizes $L(\mathbf{p})$ regardless of the values of the other parameters p_j. In the directed graph model, if node i has no outgoing arcs, then likewise we should take $p_i = 0$, and if i has no incoming arcs, then we should take $q_i = 0$.

The reparameterization we have in mind is $p_i = e^{r_i}$ and $q_i = e^{s_i}$. It is clear that the

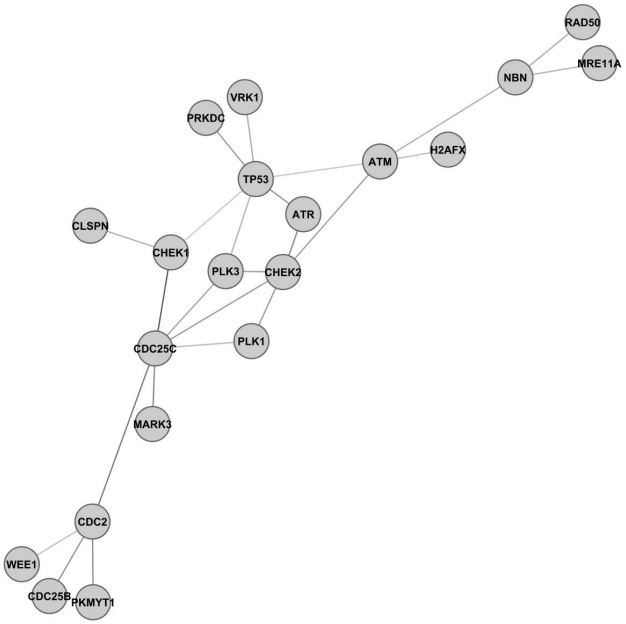


Figure 2.2: Graph of a disjoint cluster of the HPRD dataset after analysis with our method using a cutoff of ($p < 10^{-6}$). Note that this cluster is featured in the BiNGO analysis results displayed in Table 2.4.

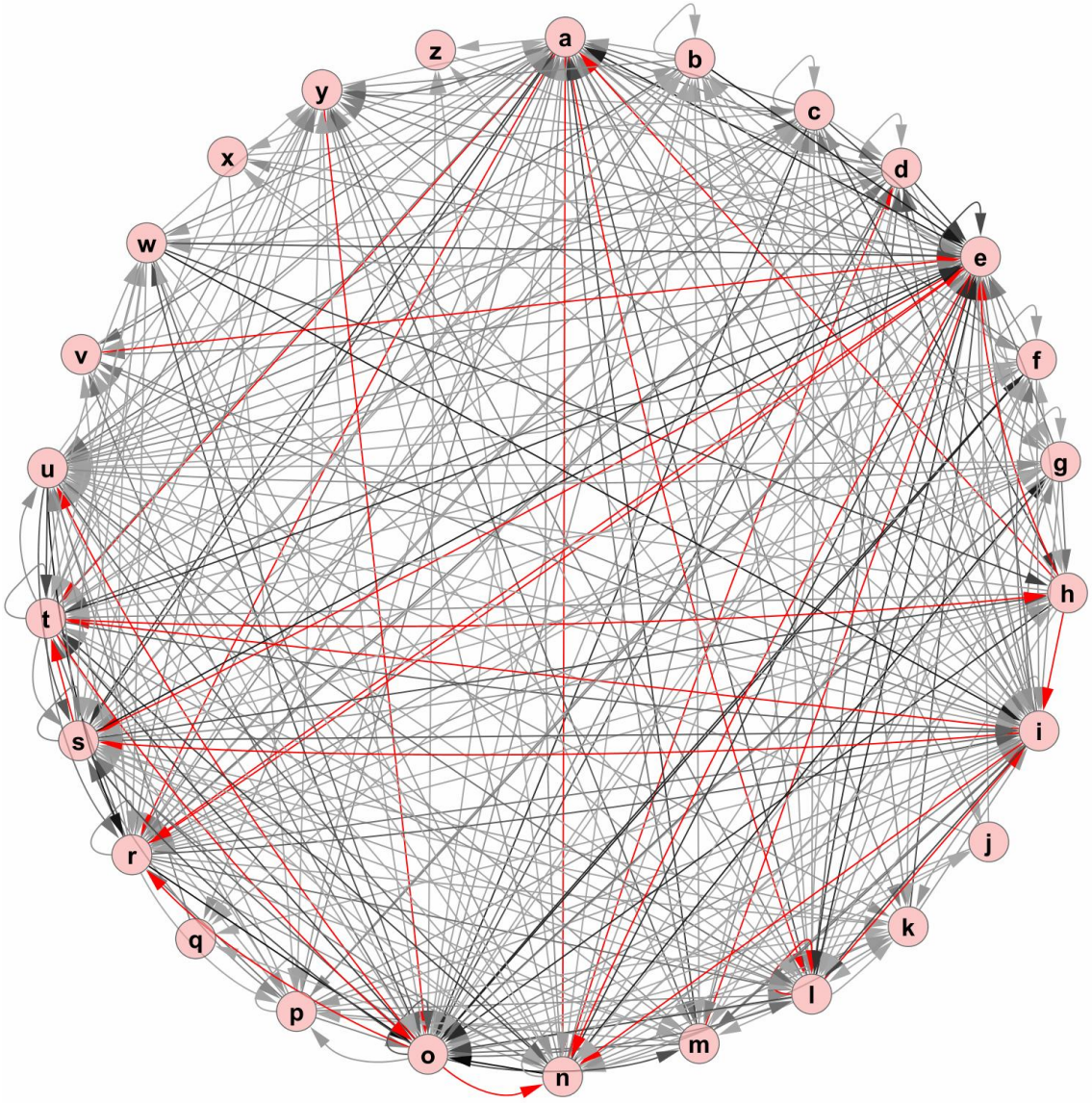


Figure 2.3: Graph of the significant connections ($p < 10^{-9}$) in the letter-pair network. In this graph, a darker edge implies a more significant connection, with the red edges highlighting the most significant connections.

reparameterized loglikelihoods

$$L(\mathbf{r}) = \sum_{\{i,j\}} [x_{ij}(r_i + r_j) - e^{r_i+r_j} - \ln x_{ij}!] \tag{2.4}$$

$$L(\mathbf{r},\mathbf{s}) = \sum_i \sum_{j\neq i} [x_{ij}(r_i + s_j) - e^{r_i+s_j} - \ln x_{ij}!] \tag{2.5}$$

are concave. If an original parameter p_i is set to 0, then we drop all terms from the loglikelihood involving r_i. If there are only two nodes, then the loglikelihood $L(\mathbf{r})$ is constant along the line $r_1 + r_2 = 0$. In the directed graph model, if an original parameter q_j is set to 0, then we drop all terms from the loglikelihood involving s_j. With only two nodes, the loglikelihood $L(\mathbf{r},\mathbf{s})$ is constant on the subspace defined by the equations $r_1 + s_2 = 0$ and $r_2 + s_1 = 0$. Strict concavity and uniqueness of the maximum likelihood estimates fail in each instance. Thus, assume that the number of nodes $m \geq 3$.

For strict concavity to hold, the positive semidefinite quadratic form

$$-v^t d^2 L(\mathbf{r})\mathbf{v} = \sum_{\{i,j\}} (v_i + v_j)^2 e^{r_i+r_j}$$

must be positive definite. When the quadratic form vanishes, $v_i + v_j = 0$ for all pairs $\{i, j\}$. If some $v_i \neq 0$, then $v_j = -v_i \neq 0$ for all $j \neq i$. With a third node k distinct from i and j, we have $v_j + v_k = -2v_i \neq 0$. This contradiction shows that $\mathbf{v} = \mathbf{0}$ and proves that $L(\mathbf{r})$ is strictly concave. It follows that there can be at most one maximum point.

In the directed graph model, it is clear that we can replace each r_i by $r_i + c$ and each s_j by $s_j - c$ without changing the value of the loglikelihood (2.4). In other words, the loglikelihood is flat along a line segment, and strict concavity fails. If we impose the constraint $r_1 = 0$ corresponding to $p_1 = 1$, then things improve. Consider the semipositive definite quadratic form

$$-\mathbf{w}^t d^2 L(\mathbf{r},\mathbf{s})\mathbf{w} = \sum_i \sum_{j\neq i} (u_i + v_j)^2 e^{r_i+s_j}$$

where $\mathbf{w}$ equals the concatenation of the vectors $\mathbf{u}$ and $\mathbf{v}$. The constraint $r_1 = 0$ corresponding to $p_1 = 1$ allows us to drop the variable u_1, and the term $(u_1 + v_j)^2 e^{r_1+s_j}$ of the quadratic form becomes $v_j^2 e^{s_j}$. In order for the quadratic form to vanish, we must have $v_j = 0$ for all j.

This in turn implies that all u_i must vanish for $i \neq 1$. Hence, $L(\mathbf{r}, \mathbf{s})$ is strictly concave under the proviso that $r_1 = 0$, and again we are entitled to conclude that at most one maximum point exists.

Existence rather than uniqueness of a maximum point depends on the property of coerciveness summarized by the requirement $\lim_{\|\mathbf{r}\| \to \infty} f(\mathbf{r}) = \infty$ for the convex function $f(\mathbf{r}) = -L(\mathbf{r})$. Equivalently, each of the sublevel sets $\{\mathbf{r} : f(\mathbf{r}) \leq c\}$ is compact. For a convex function $f(\mathbf{r})$, coerciveness is determined by the asymptotic function

$$f'_\infty(\mathbf{d}) \;=\; \sup_{t>0} \frac{f(t\mathbf{d}) - f(\mathbf{0})}{t} \;=\; \lim_{t \to \infty} \frac{f(t\mathbf{d}) - f(\mathbf{0})}{t}.$$

A necessary and sufficient condition for all sublevel sets of $f(\mathbf{r})$ to be compact is that $f'_\infty(\mathbf{d}) > 0$ for all vectors $\mathbf{d} \neq \mathbf{0}$ [UL01]. In the present circumstances,

$$f'_\infty(\mathbf{d}) \;=\; \sup_{t>0} \sum_{\{i,j\}} \Big[\frac{e^{t(d_i+d_j)} - 1}{t} - x_{ij}(d_i + d_j) \Big].$$

If any sum $d_i + d_j > 0$, then it is obvious that $f'_\infty(\mathbf{d}) > 0$. Thus, we may assume that all pairs satisfy $d_i + d_j \leq 0$. With this assumption in place, if some $x_{ij} > 0$, then the assumption $d_i + d_j < 0$ also gives $f'_\infty(\mathbf{d}) > 0$. Hence, we may also assume that $d_i + d_j = 0$ for all pairs with $x_{ij} > 0$. If all $d_j \leq 0$, suppose $d_i < 0$. Then there is at least one j with $x_{ij} > 0$. But this entails $d_i + d_j = 0$ and contradicts our assumption that $d_j \leq 0$. Finally, let us assume some $d_i > 0$. Then $d_j < 0$ for all $j \neq i$. If $x_{jk} > 0$ for a pair $\{j, k\}$ with $j \neq i$ and $k \neq i$, then $d_j + d_k = 0$ and either d_j or d_k is positive. This is a contradiction. Hence, all edges involve i. Because all nodes lacking edges are omitted from consideration, all nodes are connected to i. In other words, the only way the condition $f'_\infty(\mathbf{d}) > 0$ can occur with $\mathbf{d} \neq \mathbf{0}$ is for i to serve as a hub in the narrow sense of attracting all edges.

A hub formation is incompatible with coerciveness. Indeed, suppose i is the hub. If we take $r_i = t > 0$ and all $r_j = -t$ for $j \neq i$, then the loglikelihood (2.4) becomes

$$L(\mathbf{r}) \;=\; \sum_{j \neq i} [x_{ij}(t - t) - e^{t-t} - \ln x_{ij}!] - \sum_{\{j,k\}: j \neq i, k \neq i} e^{-2t},$$

which is bounded below as $t \to \infty$. A two-node model obviously involves two hubs.

Hubs also supply the only exceptions to coerciveness in the directed graph model. In proving this assertion, we let I be the set of nodes with incoming arcs and O be the set of nodes with outgoing arcs. The parameter r_i is defined provided $i \in O$, and the parameter s_j is defined provided $j \in I$. Suppose i is a hub with both outgoing and incoming arcs. Set $r_i = 0$, $s_i = t$, $s_j = 0$ when $j \in I \setminus \{i\}$, and $r_j = -t$ when $j \in O \setminus \{i\}$. The loglikelihood

$$\begin{aligned} L(\mathbf{r}, \mathbf{s}) &= \sum_{j \in I \setminus \{i\}} [x_{ij} 0 - e^0 - \ln x_{ij}!] \\ &\quad + \sum_{j \in O \setminus \{i\}} [x_{ji}(-t + t) - e^{-t+t} - \ln x_{ij}!] \\ &\quad - \sum_{j \in O \setminus \{i\}} \sum_{k \in I \setminus \{i,j\}} e^{-t} \end{aligned}$$

remains bounded as t tends to ∞. Thus, $L(\mathbf{r}, \mathbf{s})$ fails to be coercive in this setting.

In proving the converse for a directed graph, we write the asymptotic function as

$$f'_\infty(\mathbf{c}, \mathbf{d}) = \sup_{t>0} \sum_{i \in O} \sum_{j \in I \setminus \{i\}} \Big[\frac{e^{t(c_i + d_j)} - 1}{t} - x_{ij}(c_i + d_j) \Big].$$

A pair (i, j) is said to be active provided $i \in O$ and $j \in I$. If the loglikelihood is not coercive, then there exists a vector $(\mathbf{c}, \mathbf{d}) \neq \mathbf{0}$ with $f'_\infty(\mathbf{c}, \mathbf{d}) = 0$, where $\mathbf{c}$ is the vector of defined c_i and d is the vector of defined d_j. It suffices to show that $f'_\infty(\mathbf{c}, \mathbf{d}) = 0$ for some nontrivial $(\mathbf{c}, \mathbf{d})$ is impossible unless the graph is organized as a hub with both incoming and outgoing arcs.

Without loss of generality, we can assume that $x_{12} > 0$; otherwise, we relabel the nodes so that some arc starts at node 1 and ends at node 2. This choice also allows us to eliminate the propensity r_1 and set $c_1 = 0$. If $c_i + d_j > 0$ for an active pair (i, j), then it is obvious that $f'_\infty(\mathbf{c}, \mathbf{d}) > 0$. Furthermore, if $x_{ij} > 0$ and $c_i + d_j < 0$, then we also have $f'_\infty(\mathbf{d}) > 0$. Thus, we may assume that all active pairs (i, j) satisfy $c_i + d_j \leq 0$, with equality when $x_{ij} > 0$. Given these restrictions, the assumption $c_1 = 0$ requires that $d_j \leq 0$ for all $j \neq 1$ in I. In view of our assumption $x_{12} > 0$, we find that $d_2 = 0$. If $k \neq 2$ is in O, the restriction $c_k + d_2 \leq 0$ implies that $c_k \leq 0$. Thus, the only two components that can be positive are d_1 and c_2. Suppose the pair $(2, 1)$ is active. The inequality $c_2 + d_1 \leq 0$ implies that if either component d_1 or c_2 is positive, then the other component is negative. Similarly, if $x_{kl} > 0$

for nodes $k \neq 2$ and $l \neq 1$, then the equality $c_k + d_l = 0$ and the nonpositivity of c_k and d_l yield $c_k = d_l = 0$.

If we can show that c_2 and d_1 are nonpositive when defined, then all components of $(\mathbf{c}, \mathbf{d})$ will be nonpositive. This state of affairs actually implies that all components are 0, contradicting our assumption that $(\mathbf{c}, \mathbf{d})$ is nontrivial. To prove this claim, consider a defined component c_i. Because there exists a node j with $x_{ij} > 0$, the equation $c_i + d_j = 0$ entails $c_i = 0$ when all components of $(\mathbf{c}, \mathbf{d})$ are nonpositive. Likewise, for every defined d_j, there exists a node i with $x_{ij} > 0$. The equation $c_i + d_j = 0$ now entails $d_j = 0$ when all components of $(\mathbf{c}, \mathbf{d})$ are nonpositive.

The proof now separates into cases. In the first case, no other arcs impinge on node 1 or node 2 except possibly the arc $2 \to 1$. If the arc $2 \to 1$ does not exist, d_1 and c_2 are undefined, and we are done. If $2 \to 1$ exists, then to a avoid a hub with both incoming and outgoing arcs, there must be a third arc $k \to l$ distinct from $1 \to 2$ and $2 \to 1$. We have already observed that $c_k = d_l = 0$ for an arc $k \to l$ with $k \neq 2$ and $l \neq 1$. Therefore, the requirement $c_k + d_1 \leq 0$ entails $d_1 \leq 0$. Similarly, the requirement $c_2 + d_l \leq 0$ entails $c_2 \leq 0$.

In the second case, component d_1 is defined and component c_2 is undefined. To prevent node 1 from being a hub with both incoming and outgoing arcs, there must be an arc $k \to l$ with k and l different from 1. Because c_2 is undefined, $k \neq 2$. Hence, again $c_k = d_l = 0$. The requirement $c_k + d_1 \leq 0$ now implies $d_1 \leq 0$.

In the third case, component d_1 is undefined and component c_2 is defined. To prevent node 2 from being a hub with both incoming and outgoing arcs, there must be an arc $k \to l$ with k and l different from 2. Because d_1 is undefined, $l \neq 1$. Hence, again $c_k = d_l = 0$. The requirement $c_2 + d_l \leq 0$ now implies $c_2 \leq 0$.

In the fourth and final case, both components d_1 and c_2 are defined. The hub hypothesis fails if there exists an arc $k \to l$ with k and l both differing from 1 and 2. As noted earlier, this leads to the conclusions $d_1 \leq 0$ and $c_2 \leq 0$. If no such arc exists, then consider arcs $k \to 1$ and $2 \to l$. If the only possible k is $k = 2$, then node 2 is a hub with both incoming and outgoing arcs. Assuming $k \neq 2$, we have $c_k \leq 0$. The requirement $c_k + d_1 = 0$ now

implies $d_1 \geq 0$. In similar fashion, if the only possible value of l is 1, then node 1 is a hub with both incoming and outgoing arcs. Assuming $l \neq 1$, we have $d_l \leq 0$. The requirement $c_2 + d_l = 0$ now implies $c_2 \geq 0$. Unless $d_1 = c_2 = 0$, the two conditions $d_1 \geq 0$ and $c_2 \geq 0$ are incompatible with our earlier finding that $d_1 > 0$ implies $c_2 < 0$ and vice versa.

In summary, we have found that the condition $f'_\infty(\mathbf{c}, \mathbf{d}) = 0$ and the assumption of no hub with both incoming and outgoing arcs imply that $(\mathbf{c}, \mathbf{d}) = \mathbf{0}$. Thus, the strictly convex function $f(\mathbf{r}, \mathbf{s}) = -L(\mathbf{r}, \mathbf{s})$ is coercive under the no hub assumption and attains its maximum at a unique point.

2.8.2 Convergence of the MM Algorithms

Verification of global convergence of the MM algorithms hinge on five properties of the objective function $L(\mathbf{p})$ and the iteration map $M(\mathbf{p})$:

(a) $L(\mathbf{p})$ is coercive,

(b) $L(\mathbf{p})$ has only isolated stationary points,

(c) $M(\mathbf{p})$ is continuous,

(d) A point is a fixed point of $M(\mathbf{p})$ if and only if it is a stationary point of $L(\mathbf{p})$,

(e) $L[M(\mathbf{p})] \leq L(\mathbf{p})$, with equality if and only if $\mathbf{p}$ is a fixed point of $M(\mathbf{p})$.

See the reference [Lan04] for full details.

Verification of these properties in the multigraph models is straightforward. Coerciveness has already been dealt with under the reparameterization $p_i = e^{r_i}$ and the no hub assumption. Because the reparameterized loglikelihood $L(\mathbf{r})$ is strictly concave, there is a single stationary point in both the original and transformed coordinates. Inspection of the iteration map (Equation 3 in paper) shows that it is continuous. It does involve a division by a denominator that could tend to 0, but this contingency is ruled out by coerciveness. The fixed point condition $M(\mathbf{p}) = \mathbf{p}$ occurs when the surrogate function satisfies the equation $\nabla g(\mathbf{p} \mid \mathbf{p}) = \mathbf{0}$. The identity $\nabla L(\mathbf{p}) = \nabla g(\mathbf{p} \mid \mathbf{p})$ at every interior point of the domain of the objective

function shows that fixed points and stationary points coincide. Finally, the strict concavity of the surrogate function $g(\mathbf{p} \mid \mathbf{p}^n)$ demonstrates that $g(\mathbf{p}^{n+1} \mid \mathbf{p}^n)$ is strictly larger than $g(\mathbf{p}^n \mid \mathbf{p}^n)$ unless $\mathbf{p}^{n+1} = \mathbf{p}^n$. Because $g(\mathbf{p} \mid \mathbf{p}^n)$ minorizes $L(\mathbf{p})$, this ascent property carries over to $L(\mathbf{p})$. With minor notational changes, the same arguments apply to the directed graph model.

2.8.3 Log P-Value Approximations

Since the extreme right-tail probabilities of the Poisson distribution lead to computer underflows, we must resort to approximation. Let the Poisson random deviate X have mean λ. For n much larger than λ, we find that

$$\begin{aligned}
\Pr(X \geq n) &= \sum_{k=n}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \frac{e^{-\lambda}\lambda^n}{n!} \sum_{k=0}^{\infty} \frac{\lambda^k n!}{(n+k)!} \\
&\leq \frac{e^{-\lambda}\lambda^n}{n!} \sum_{k=0}^{\infty} \left(\frac{\lambda}{n}\right)^k \\
&= \frac{e^{-\lambda}\lambda^n}{n!} \left(\frac{1}{1-\frac{\lambda}{n}}\right) \\
&= \frac{e^{-\lambda}\lambda^n}{(n-1)!} \frac{1}{n-\lambda}.
\end{aligned}$$

Because n is large, we can approximate $(n-1)!$ by Stirling's formula

$$(n-1)! \approx \sqrt{2\pi} n^{n-1/2} e^{-n}.$$

This allows us to take logarithms of $\Pr(X \geq n) \approx \frac{e^{n-\lambda}\lambda^n}{\sqrt{2\pi} n^{n-1/2}(n-\lambda)}$ in the construction of our tables.

2.8.4 Appendix Tables and Figures

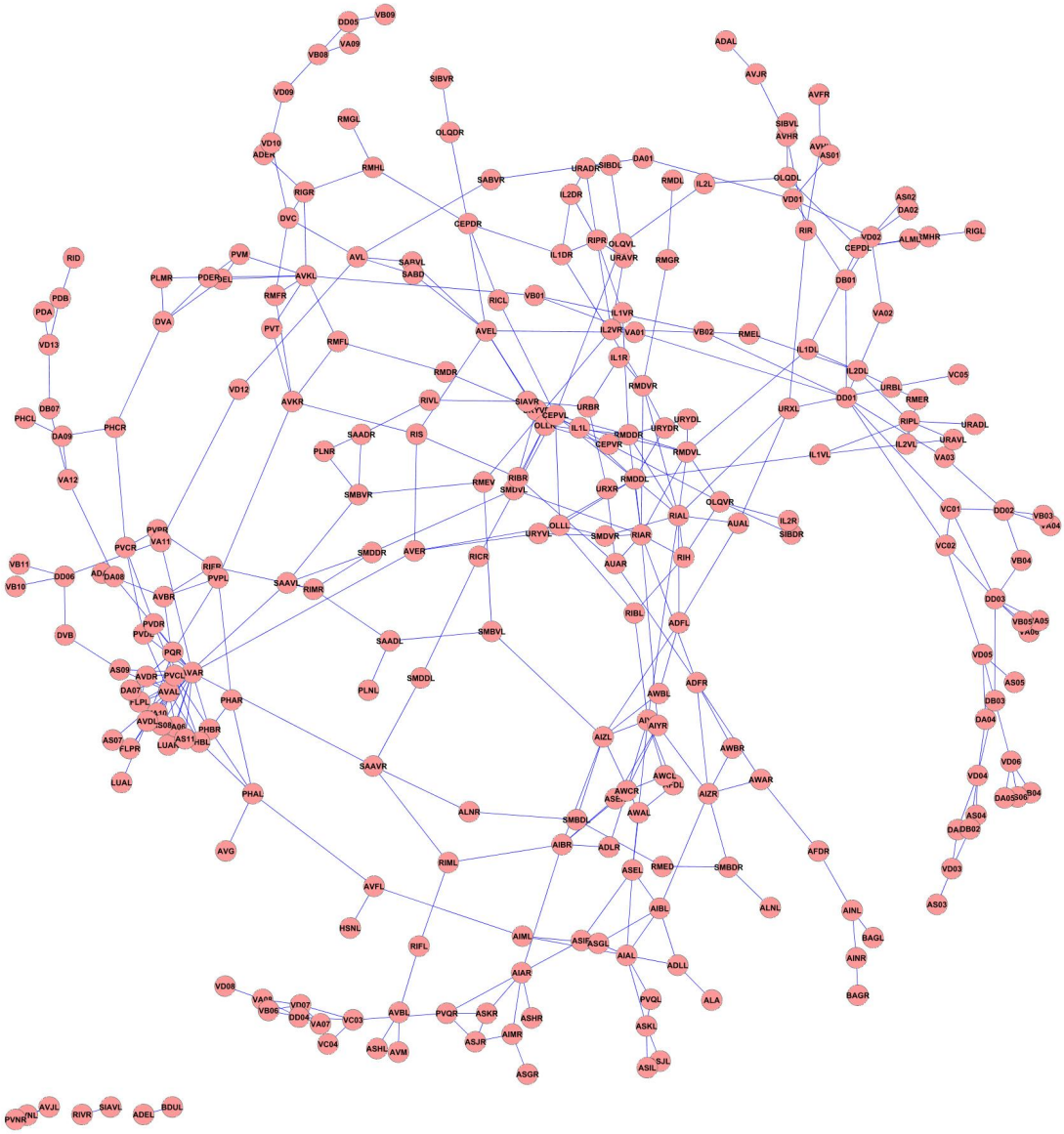


Figure 2.4: Graph of *C. Elegans* neural network with a p-value of 10^{-6}.

Table 2.1: List of the 20 most significant connections of the *C. elegans* dataset. To the right of each pair appear the observed number of edges, the expected number of edges, and minus the log base 10 p-value.

RANK	**NEURON1**	**NEURON2**	**OBS.**	**EXP.**	**-LOGP**
1	VB03	DD02	37	0.7967	47.1265
2	VB08	DD05	30	0.382	45.1218
3	VB06	DD04	30	0.4653	42.5846
4	VB05	DD03	27	0.6609	33.1679
5	VD03	DA03	24	0.5834	29.6503
6	VA06	DD03	24	0.6495	28.5599
7	VA08	DD04	21	0.4289	27.6046
8	VD05	DB03	23	0.6934	26.3561
9	VA04	DD02	21	0.6325	24.1455
10	PDER	AVKL	16	0.2738	22.4316
11	VB02	DD01	20	0.6488	22.4101
12	RIPR	IL2VR	14	0.1702	21.7724
13	VA09	DD05	15	0.2934	20.2217
14	PDER	DVA	16	0.3972	19.8949
15	OLLL	AVER	18	0.6434	19.5152
16	VD03	AS03	14	0.2599	19.2348
17	VD03	DB02	16	0.4868	18.5184
18	VD01	DA01	14	0.3102	18.1794
19	RIPL	IL2VL	11	0.1136	18.0317
20	VA03	DD01	18	0.7851	18.0170

Table 2.2: Top 20 proteins with the most observed connections in the literature curated protein database.

RANK	PROTEIN	OBS.	SIG.	PROP.
1	TP53	358	6	1.2515
2	GRB2	291	3	1.0164
3	SRC	277	5	0.9674
4	YWHAG	249	0	0.8693
5	CREBBP	231	0	0.8063
6	EGFR	231	5	0.8063
7	EP300	231	0	0.8063
8	PRKCA	229	4	0.7993
9	MAPK1	213	4	0.7433
10	CSNK2A1	207	1	0.7223
11	FYN	205	4	0.7153
12	PRKACA	202	2	0.7048
13	ESR1	200	1	0.6978
14	SHC1	195	5	0.6803
15	SMAD3	193	0	0.6733
16	STAT3	190	10	0.6628
17	SMAD2	183	1	0.6384
18	RB1	169	2	0.5894
19	TRAF2	168	2	0.5859
20	SMAD4	166	0	0.5789

Table 2.3: The 20 proteins with the most significant connections ($p < 10^{-6}$) in the literature curated protein database.

RANK	**PROTEIN**	**OBS.**	**SIG.**	**PROP.**
1	STAT3	190	10	0.6628
2	STAT1	162	9	0.565
3	MAPT	127	9	0.4427
4	PCNA	114	8	0.3973
5	RPS6KA1	59	7	0.2055
6	TP53	358	6	1.2515
7	MAPK3	148	6	0.5161
8	PTPN6	144	6	0.5021
9	DLG4	132	6	0.4602
10	MAPK14	107	6	0.3729
11	BTK	100	6	0.3485
12	HCK	82	6	0.2857
13	CREB1	59	6	0.2055
14	CDC25C	58	6	0.202
15	F2	57	6	0.1985
16	COPS4	31	6	0.1079
17	SRC	277	5	0.9674
18	EGFR	231	5	0.8063
19	SHC1	195	5	0.6803
20	LCK	156	5	0.544

Table 2.4: BiNGO results of the small detached component around TP53 (Figure 2.2) in the literature curated protein database [MHK05]. Note here that the p-values reported in the column labeled -LOG P are the BiNGO p-values for clustering, not the p-values delivered by the Poisson model.

GO-ID	**-LOGP**	**GO TERM**
7049	15.8761	cell cycle
6974	12.6819	response to DNA damage stimulus
279	12.2596	M phase
6281	12.1261	DNA repair
22403	11.5544	cell cycle phase
22402	11.5421	cell cycle process
6259	11.4597	DNA metabolic process
43283	9.3883	biopolymer metabolic process
43687	8.9393	post-translational protein modification
6796	8.2857	phosphate metabolic process
6793	8.2857	phosphorus metabolic process
7126	8.0123	meiosis
51327	8.0123	M phase of meiotic cell cycle
51321	7.9706	meiotic cell cycle
6464	7.6440	protein modification process
6302	7.6216	double-strand break repair
6310	7.5607	DNA recombination
43170	7.5607	macromolecule metabolic process
43412	7.5186	biopolymer modification
6468	7.5171	protein amino acid phosphorylation
74	7.4559	regulation of cell cycle
42770	7.3665	DNA damage response, signal transduction

Table 2.5: Most significantly connected word pairs.

RANK	-LOGP	OBS.	EXPECTED	PAIR
1	391.3236	355	10.7509	i am
2	332.9314	293	8.2031	my lord
3	220.4243	337	30.4288	i have
4	195.8137	286	23.9518	i will
5	173.4930	73	0.1179	lady macbeth
6	163.1923	105	1.1239	thou art
7	160.2825	215	15.5290	it is
8	159.2199	399	70.5448	in the
9	146.6971	111	2.0425	no more
10	128.5489	51	0.0600	re enter
11	124.9406	160	10.6422	i know
12	110.9513	109	4.1161	let me
13	107.6928	151	11.8937	you are
14	107.3818	66	0.6054	second lord
15	95.2465	168	19.1548	i do
16	94.4514	80	2.0708	they are
17	94.0240	83	2.4030	pray you
18	93.8222	61	0.6902	thou hast
19	93.6175	137	11.6537	i would
20	88.9511	43	0.1446	first soldier

Table 2.6: Words observed as a pair and never as singletons.

PAIR	**PAIR**
hysterica passio	ordered honorably
bosko chimurcho	stinkingly depending
oscorbidulchos volivorco	facit monachum
boblibindo chicurmurco	stench consumption
suit's unprofitable	rustic revelry
quietly debated	fellowships accurst
tu brute	du vinaigre
ovid's metamorphoses	nec arcu
sectary astronomical	penthouse lid
boarish fangs	sun's uprise
curvets unseasonably	remained unscorched
cullionly barbermonger	clothier's yard
aves vehement	parallels nessus
downfallen birthdom	et tu
threateningly replies	mort du
tick tack	kerely bonto
kneaded clod	whoop jug
brethren's obsequies	fa sol
revania dulche	mastiff greyhound
tempestuous gusts	throca movousus

Table 2.7: Most significantly connected letter pairs.

PAIR	-LOGP	OBS.	EXP.
th	10042	20308	2739
ou	3444	10452	2230
nd	3358	8125	1366
ll	2747	5404	703
yo	2257	4488	592
he	2098	15227	6085
ng	1974	3790	477
an	1775	10554	3769
ve	1717	5138	1082
in	1469	8825	3172
ow	1365	3113	489
er	1283	10264	4312
of	1186	3273	636
ha	1167	7665	2902
st	1069	5555	1823
my	999	2221	339
wi	835	3336	907
us	825	4134	1324
is	821	6346	2622
wh	778	3127	854
hi	692	5924	2573
ma	672	3585	1198
ur	659	4331	1641
fo	640	2855	843
om	619	2896	886

Table 2.8: Convergence results for each of the 5 real datasets. Note that convergence was defined as a change in loglikelihood of less than 10^{-8} percent of the previous loglikelihood. Time given is for a dual processor computer running at 2.4 GHz. Time is given in seconds (s).

Dataset	**# Nodes**	**# Edges**	**# Iterations**	**Time(s)**
Letter Pairs	27	503,951	21	42
C. Elegans	281	6,417	23	9
Protein Ints.	9,213	88,456	18	741
Word Pairs	10,789	137,338	24	1,415
Rad. Hybrid	20,145	825,551,643	29	14,903

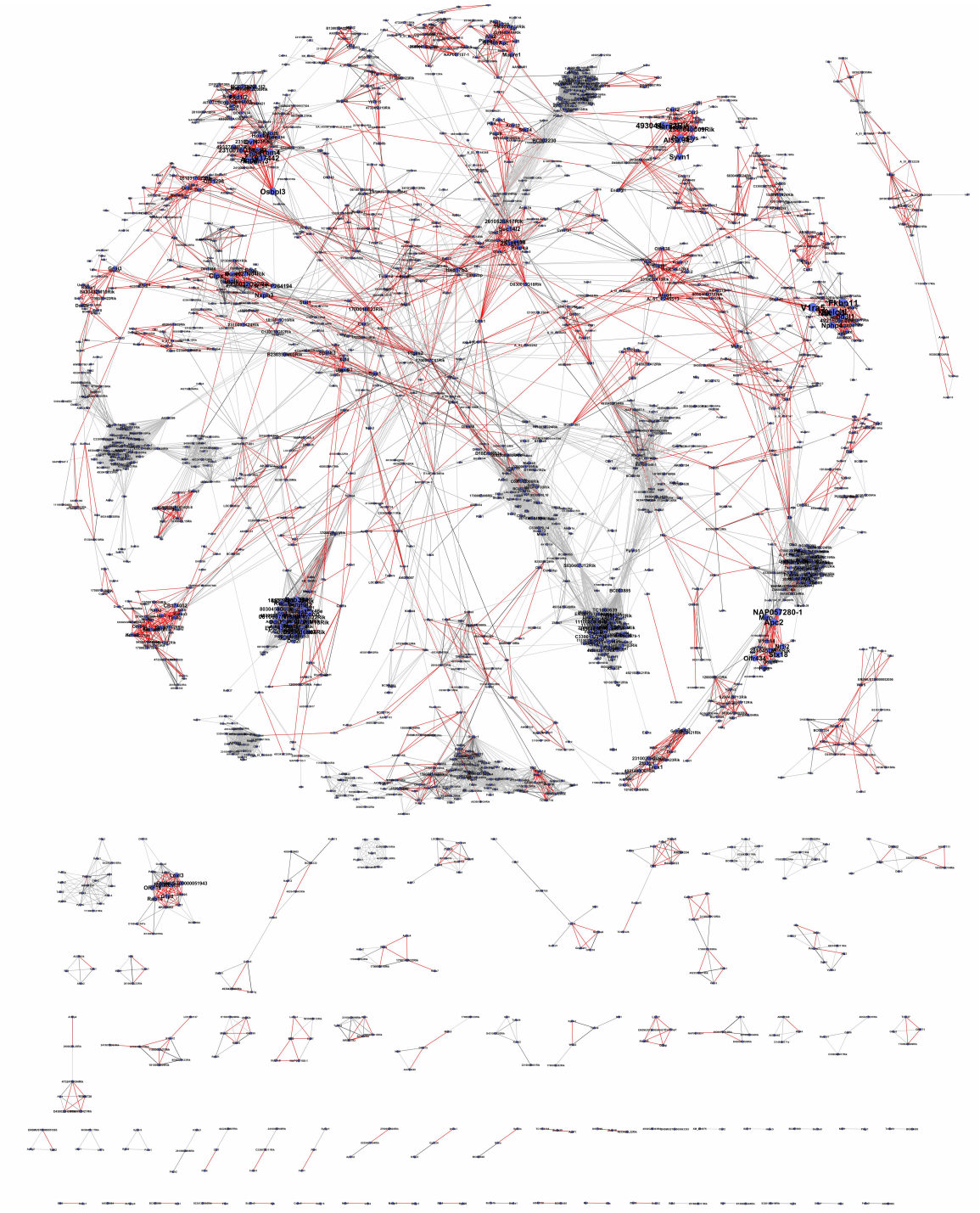


Figure 2.5: Graph of the Radiation Hybrid network. In this graph, node size is proportional to a node's estimated propensity. Also, the darker the edge, the more significant the connection; red lines highlight the most significant connections.

CHAPTER 3

Cluster and Propensity Based Approximation of a Network

3.1 Abstract

Background: The models in this article generalize current models for both correlation networks and multigraph networks. Correlation networks are widely applied in genomics research. In contrast to general networks, it is straightforward to test the statistical significance of an edge in a correlation network. It is also easy to decompose the underlying correlation matrix and generate informative network statistics such as the module eigenvector. However, correlation networks only capture the connections between numeric variables. An open question is whether one can find suitable decompositions of the similarity measures employed in constructing general networks. Multigraph networks are attractive because they support likelihood based inference. Unfortunately, it is unclear how to adjust current statistical methods to detect the clusters inherent in many data sets.

Results: Here we present an intuitive and parsimonious parametrization of a general similarity measure such as a network adjacency matrix. The cluster and propensity based approximation (CPBA) of a network not only generalizes correlation network methods but also multigraph methods. In particular, it gives rise to a novel and more realistic multigraph model that accounts for clustering and provides likelihood based tests for assessing the significance of an edge after controlling for clustering. We present a novel Majorization-Minimization (MM) algorithm for estimating the parameters of the CPBA. To illustrate the practical utility of the CPBA of a network, we apply it to gene expression data and to a bipartite network model for diseases and disease genes from the Online Mendelian Inheritance

in Man (OMIM).

Conclusions: The CPBA of a network is theoretically appealing since a) it generalizes correlation and multigraph network methods, b) it improves likelihood based significance tests for edge counts, c) it directly models higher-order relationships between clusters, and d) it suggests novel clustering algorithms. The CPBA of a network is implemented in Fortran 95 and bundled in the freely available R package PropClust.

3.1.1 Keywords

Network decomposition, model-based clustering, MM algorithm, propensity, network conformity

3.2 Background

The research of this article was originally motivated by two types of network models: correlation networks and multigraphs. After reviewing these special network models, we describe how structural insights gained from them can be used to tackle research questions arising in the study of general networks specified by network adjacencies and more generally to unsupervised learning scenarios modeled by similarity measures.

3.2.1 Background: adjacency matrix and multigraphs

Networks are used to describe the pairwise relationships between n nodes (or vertices). For example, we use networks to describe the functional relationships between n genes. We consider networks that are fully specified by an $n \times n$ **adjacency matrix** $\boldsymbol{A} = (A_{ij})$, whose entry A_{ij} quantifies the connection strength from node i to node j. For an **unweighted** network, A_{ij} equals 1 or 0, depending on whether a connection (or link or edge) exists from node i to node j.

For a **weighted network**, A_{ij} equals a real number between 0 and 1 specifying the connection strength from node i to node j. For an undirected network, the connection

strength A_{ij} from i to j equals the connection strength A_{ji} from j to i. In other words, the adjacency matrix $\boldsymbol{A}$ is symmetric. For a directed network, the adjacency matrix is typically not symmetric. Unless we explicitly mention otherwise, we will deal with undirected networks. In this paper the diagonal entries A_{ii} of the adjacency matrix $\boldsymbol{A}$ have no special meaning. We arbitrarily set them equal to 1 (the maximum adjacency value); other authors set them equal to 0 [Lux07].

In an **(unweighted) multigraph**, the adjacencies $A_{ij} = n_{ij}$ are integers specifying the number of edges between two nodes. A general similarity matrix (whose entries are non-negative real numbers possibly outside [0,1]) can be interpreted as a **weighted multigraph**. In each of the network types, the connectivities

$$k_i \;=\; \sum_{j \neq i} A_{ij} \tag{3.1}$$

are important statistics pertinent to finding highly connected hubs. In an unweighted network (a graph), k_i is the degree of node i.

3.2.2 Background: correlation- and co-expression networks

Network methods are frequently used to analyze experiments recording levels of transcribed messenger RNA. The gene expression profiles collected across samples can be highly correlated and form modules (clusters) corresponding to protein complexes, organelles, cell types, and so forth [ESB98, SSK03, OKI08]. It is natural to describe these pairwise correlations in network language. The intense interest in co-expression networks has elicited a number of new models and statistical methods for data analysis [SSK03, ZH05, HLH07, HZC06, CZF06, OHG06, CES07, KCW08], with recent applied research focusing on differential network analysis and regulatory dysfunction [DYK12, Fue10].

A correlation network is a network whose adjacency matrix $\boldsymbol{A} = (A_{ij})$ is constructed from the correlations between quantitative measurements summarized in an $m \times n$ data matrix $\boldsymbol{X} = (x_{ij})$. The m rows of $\boldsymbol{X}$ correspond to sample measurements (subjects), and the n columns of $\boldsymbol{X}$ correspond to network nodes (genes). The jth column $\boldsymbol{x}_j$ of $\boldsymbol{X}$ serves as a **node profile** across the m samples. A correlation network adjacency matrix is constructed

from the pairwise correlations $\mathrm{Corr}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in either of two ways. An unweighted gene co-expression network is defined by thresholding the absolute values of the correlation matrix. A weighted adjacency matrix is a continuous transformation of the correlation matrix. For reasons explained in [ZH05, HZC06], it is advantageous to define the adjacency A_{ij} between two genes i and j as a power $\beta \geq 1$ of the absolute value of their correlation coefficient; thus,

$$A_{ij} \;=\; |\,\mathrm{Corr}(\boldsymbol{x}_i, \boldsymbol{x}_j)|^{\beta}.$$

Weighted gene co-expression networks have found many important medical applications, including identifying brain cancer genes [HZC06], characterizing obesity genes [GDZ06, FGA07], understanding atherosclerosis [GGC06], and locating the differences between human and chimpanzee brains [OHG06]. One of the important steps of weighted correlation network analysis is to find network modules, usually via hierarchical clustering. Each module (cluster) is then represented by the module eigengene defined by a certain singular value decomposition (SVD). Suppose $\boldsymbol{Y}$ denotes the expression data of a single module (cluster) after the appropriate columns of $\boldsymbol{X}$ have been extracted and standardized to have mean 0 and variance 1. The SVD of $\boldsymbol{Y}$ is the decomposition $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^t$, where the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal, $\boldsymbol{D}$ is a diagonal matrix with nonnegative diagonal entries (singular values) presented in descending order, and the superscript t indicates a matrix or vector transpose. The sign of the dominant singular vector $\boldsymbol{u}_1$ (the first column of $\boldsymbol{U}$) is fixed by requiring a positive average correlation with the columns of $\boldsymbol{Y}$; $\boldsymbol{u}_1$ is referred to as the module eigenvector or eigengene. One can show that $\boldsymbol{u}_1$ is an eigenvector of the $m \times m$ sample correlation matrix $\frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^t$ corresponding to the largest eigenvalue. The eigenvector $\boldsymbol{u}_1$ explains the maximum amount of variation in the columns of $\boldsymbol{Y}$.

Let d_i be the ith singular value of $\boldsymbol{Y}$. The eigenvector factorizability

$$\mathrm{EF}(\boldsymbol{u}_1) \;=\; \frac{|d_1|^4}{\sum_j |d_j|^4}$$

measures how well a network factors [HD08]. This measure is very similar to the proportion of variation explained, $d_1^2 / \sum d_j^2$. One can prove [HD08] that when $\mathrm{EF}(u_1) \approx 1$, the correlation matrix $\boldsymbol{Y}$ approximately factors as

$$\mathrm{Corr}(\boldsymbol{x}_i, \boldsymbol{x}_j) \;\approx\; \mathrm{Corr}(\boldsymbol{x}_i, \boldsymbol{u}_1)\,\mathrm{Corr}(\boldsymbol{x}_j, \boldsymbol{u}_1)\,.$$

In co-expression networks, modules are often approximately factorizable [DH07, HD08]. For a network comprised of multiple modules, it should come as no surprise that when the eigenvector factorizabilities of all modules are close to 1, the correlation network factors as

$$A_{ij} \approx |\operatorname{Corr}(\boldsymbol{x}_i, \boldsymbol{u}_1^{c_i})|^\beta |\operatorname{Corr}(\boldsymbol{x}_j, \boldsymbol{u}_1^{c_j})|^\beta |\operatorname{Corr}(\boldsymbol{u}_1^{c_i}, \boldsymbol{u}_1^{c_j})|^\beta \approx p_i p_j r_{c_i c_j}, \tag{3.2}$$

where $\boldsymbol{u}_1^{c_i}$ is the module eigenvector of the module containing i, $p_i = |\operatorname{Corr}(\boldsymbol{x}_i, \boldsymbol{u}_1^{c_i})|^\beta$ measures the intramodular connectivity (module membership) of node i with respect to its module, and $r_{c_i c_j} = |\operatorname{Corr}(\boldsymbol{u}_1^{c_i}, \boldsymbol{u}_1^{c_j})|^\beta$ measures the similarity between clusters c_i and c_j. The quantity

$$\mathrm{kME}_i = \operatorname{Corr}(\boldsymbol{x}_i, \boldsymbol{u}_1^{c_i}) \tag{3.3}$$

is called the module membership measure or conformity [DH07, HD08].

Unlike general networks, correlation networks allow assessment of the statistical significance of an edge (via a correlation test) and generate informative network statistics such as the module eigenvector. But correlation network methods can only be applied to model the correlations between numeric variables. An open question is whether correlation network methods can be generalized to general networks by defining a suitable decomposition of a general network similarity measure. In the following, we will address this question.

3.3 Results and discussion

3.3.1 CPBA is a sparse approximation of a similarity measure

Consider a general $n \times n$ symmetric adjacency matrix A, for example one generated by a multigraph. Because the diagonal entries of $\boldsymbol{A}$ are irrelevant, $\boldsymbol{A}$ is determined by its $\binom{n}{2}$ upper-diagonal entries. We now describe a low-rank matrix approximation to $\boldsymbol{A}$ based on partitioning the n nodes into K clusters labeled $1, \ldots, K$. Motivated by (Eq. 3.2), our approximation of a general similarity relies on three main ingredients. The first is a **cluster assignment** indicator $\boldsymbol{c} = (c_i)$ whose entry c_i equals a when node i belongs to cluster a. The cluster label $a = 0$ is special and is reserved for singleton nodes outside any of the "proper"

clusters. The clusters are required to be non-empty except for the improper cluster 0. The second ingredient is a $K \times K$ **cluster similarity matrix** $\boldsymbol{R} = (r_{ab})$ whose entries quantify the relationships between clusters. The third and final ingredient is the **propensity vector** $\boldsymbol{p} = (p_i)$ whose components quantify the tendency (propensity) of the various nodes to form edges. The goal of cluster and propensity based approximation (CPBA) is to construct an approximation to $\boldsymbol{A}$ by optimally choosing the cluster assignment indicator $\boldsymbol{c}$, the cluster similarity matrix $\boldsymbol{R}$, and the propensity vector $\boldsymbol{p}$. CPBA assumes that the adjacency matrix A_{ij} can be approximated as

$$A_{ij} \approx r_{c_i c_j} p_i p_j \,. \tag{3.4}$$

The right-hand side with $\binom{K}{2} + n$ parameters can be interpreted as a sparse parametrization of the left-hand side with $\binom{n}{2}$ parameters. In a weighted correlation network, the propensity p_i of node i is approximately $|\text{kME}_i|^\beta$. The cluster similarity r_{ab}, defined by the correlation $|\operatorname{Corr}(\boldsymbol{u}_1^a, \boldsymbol{u}_1^b)|^\beta$ between eigengenes, is an intuitive measure of the interactions between modules. The diagonal entries r_{aa} of $\boldsymbol{R}$ are identically 1.

3.3.2 Objective functions for estimating CPBA

In practice, CPBA parameters $\boldsymbol{c}$, $\boldsymbol{p}$, and $\boldsymbol{R}$ of a general similarity are unknown and must be estimated by optimizing a suitably defined objective function. In this article, we describe estimation methods that are based on optimizing two superficially different objective functions. Our first objective is just the squared Frobenius matrix norm

$$\text{Frobenius}(\boldsymbol{c}, \boldsymbol{p}, \boldsymbol{R}) = \sum_i \sum_{j \neq i} (A_{ij} - r_{c_i c_j} p_i p_j)^2. \tag{3.5}$$

Our second objective is the Poisson log-likelihood

$$\begin{aligned} \text{Poisson}(\boldsymbol{c}, \boldsymbol{p}, \boldsymbol{R}) &= \sum_i \sum_{j \neq i} \ln \left[\frac{(r_{c_i c_j} p_i p_j)^{A_{ij}} e^{-r_{c_i c_j} p_i p_j}}{A_{ij}!} \right] \\ &= \sum_i \sum_{j \neq i} \left[A_{ij} \ln(r_{c_i c_j} p_i p_j) - r_{c_i c_j} p_i p_j - \ln(A_{ij}!) \right]. \end{aligned} \tag{3.6}$$

Our later multigraph example interprets $\text{Poisson}(\boldsymbol{c}, \boldsymbol{p}, \boldsymbol{R})$ in this traditional sense. The functional form of the Poisson log-likelihood even applies when the A_{ij} are non-integer. The

factorial $A_{ij}!$, which is irrelevant to maximization in any case, can then be defined via the gamma function. In practice maximization of the Poisson log-likelihood and minimization of the Frobenius norm yield very similar numerical updates.

In the Methods section, we describe a powerful MM algorithm for optimizing the objective functions and estimating its parameters. We now pause and briefly describe a few major applications. First, the sparse parametrization can be used to derive relationships between network statistics; our previous research highlights this possibility [DH07, HD08]. For example, the connectivity index (Eq. 3.1) can be approximated by

$$k_i \;=\; \sum_{j \neq i} A_{ij} \;\approx\; p_i \sum_{j \neq i} r_{c_i c_j} p_j. \tag{3.7}$$

Second, since our optimization algorithms also strive to choose the best cluster assignment indicator $\boldsymbol{c}$, they naturally give rise to clustering algorithms. Cluster reassignment is carried out node by node in a sequential fashion. For the sake of computationally efficiency, all parameters are fixed until node reassignment has stabilized. If parameters are updated as each node is visited, then the computational overhead seriously hinders analysis of networks with ten thousand nodes. Our limited experience suggests that more frequent re-estimation of parameters is less likely to end with an inferior optimal configuration. Hence, the tradeoff is complex.

Other major uses depend on the underlying model. In the Frobenius setting, the model can be used to generalize conformity-based decomposition of a network as shown in Example 2. In the Poisson log-likelihood setting, our model suggest a new clustering procedure. In contrast to other clustering procedures, the CPBA models provide a means of relating clusters to each other via the cluster similarities r_{ab}. Furthermore, likelihood based objective functions permit statistical tests for assessing the significance of an edge. For example, in the multigraph model, the significance of the number of connections between two nodes can be ascertained by comparing the observed number of connections to the expected number of connections under the Poisson model. Finally, likelihood based objective functions provide a rational basis for estimating the number of clusters in a data set.

In the following three examples, we illustrate how to generalize a variety of network

models to include clustering.

3.3.3 Example 1: Generalizing the random multigraph model

We recently explored a random multigraph model[RAS10] that allows multiple edges to form between two nodes and edges to form with different probabilities. Edges still form independently. Under the random multigraph model, each node i is assigned a propensity p_i. The random number of edges between nodes i and j is then assumed to follow a Poisson distribution with mean $p_i p_j$. This model relies entirely on propensities and ignores cluster similarities. We will refer to it as the Pure Propensity Poisson Model (PPP) to avoid confusion with CPBA. Thus, the PPP log-likelihood is

$$\begin{aligned} \text{Pure Propensity Poisson}(\boldsymbol{p}) &= \sum_i \sum_{j \neq i} \ln \left[\frac{(p_i p_j)^{A_{ij}} e^{-p_i p_j}}{A_{ij}!} \right] \\ &= \sum_i \sum_{j \neq i} \left[A_{ij} \ln(p_i p_j) - p_i p_j - \ln(A_{ij}!) \right] \\ &= \sum_i \sum_{j \neq i} \left[n_{ij} \ln(p_i p_j) - p_i p_j - \ln(n_{ij}!) \right], \end{aligned} \tag{3.8}$$

where $A_{ij} = n_{ij}$ is the number of edges between nodes i and j. While future work could explore alternatives to the Poisson distribution, it is attractive for several reasons. First, it is the simplest model that gives the requisite flexibility. Second, a Poisson random variable accurately approximates a sum of independent Bernoulli random variables. A binomial distribution also serves this purpose, but it imposes a hard upper bound on the number of successes. Third, the Poisson model is convenient mathematically since it yields nice MM updates in maximum likelihood estimation of the model parameters[RAS10]. Fourth, a likelihood formulation permits testing for statistically significant connections between nodes.

Although the parametrization (Eq. 3.8) of PPP is flexible and computationally tractable, it ignores cluster formation. To address this limitation, we propose to exploit the CPBA parametrization. This extension is natural because many large multigraphs appear to be made up of smaller sub-networks, often referred to as modules, that are highly connected internally and only sparsely connected externally. For example, consider a co-authorship multigraph where an edge is placed between two scientists whenever they co-author an

article. Scientists working at the same institution and in the same department tend to be highly connected. Similarly, scientists tend to collaborate with other scientists working on the same research topics. Cluster structure is also inherent in biology. For instance, genes often function in pathways, and proteins often cluster in evolutionary families. Thus, when a network exhibits clustering, the propensity to form connections within a cluster is usually higher than the propensity to form connections between clusters. This phenomenon cannot be modeled using our original PPP model [RAS10] and provides the motivation for injecting cluster similarity into the multigraph model. Our hope is that the CPBA based multigraph model will better account for differences in intracluster and intercluster connections and lead to better identification of significant connections. In the absence of an explicit model for clustering, the PPP model is likely to falter on a dataset that exhibits clustering. The most likely result is a host of significant connections between nodes in the same cluster since they all exhibit more edges than expected by chance. These types of significant connections are often uninteresting. In the above mentioned co-authorship network, the cluster structure may reflect institutional affiliations. In this case, it may be more interesting to identify pairs of researchers who publish more (or less) than is expected based on their workplace location.

To keep the number of parameters to a minimum, the cluster similarity matrix $\boldsymbol{R} = (r_{ab})$ is assumed to be symmetric with a unit diagonal. Thus, our new random multigraph model, CPBA, adds just $\binom{K}{2}$ parameters for K clusters. As previously postulated, the number of edges between nodes i and j in clusters c_i and c_j is Poisson distributed with mean $r_{c_i c_j} p_i p_j$.

3.3.4 Example 2: Generalizing the conformity-based decomposition of a network

To demonstrate the value in our clustering model and tap into the wealth of data on weighted networks [Hor11], we propose a clustering extension. Because weighted networks by definition have edge weights in $[0, 1]$, we drop the Poisson assumption and instead minimize the Frobenius criterion (Eq. 3.5). A major benefit of this model is that it generalizes the conformity-based decomposition of a network [Hor11]. An adjacency matrix $\boldsymbol{A} = (A_{ij})$ is

exactly factorizable if and only if there exists a vector $\boldsymbol{f} = (f_i)$ with non-negative elements such that

$$A_{ij} = f_i f_j \tag{3.9}$$

for all $i \neq j$. In this setting, f_i is often called the **conformity** of node i. Although the term **factorizable network** was first proposed in [DH07], numerous examples of these types of networks can be found in the literature. A physical model for experimentally determined protein-protein interactions is exactly factorizable [DAS06]. In that model, the **affinity** A_{ij} between proteins i and j is the product of conformities $f_i = \exp(-\alpha_i)$, where α_i is the number of hydrophobic residues on protein i. Since it can be shown that $\boldsymbol{f}$ is uniquely defined if the network contains $n \geq 3$ nodes and all $A_{ij} > 0$ [DH07, Hor11], it is easy to see that the propensity vector matches the conformity vector, $\boldsymbol{p} = \boldsymbol{f}$, when all $r_{ab} = 1$. Even when a network is not factorizable, our method can estimate conformities while simultaneously clustering the nodes into more factorizable modules. In addition, the entries of the cluster similarity matrix $\mathbf{R}$ can be interpreted as adjacencies between modules. Thus, the cluster similarity matrix represents a network whose nodes are networks themselves. In correlation network applications, we proposed a similar measure [LH07], and for gene networks we defined a measure of the probability of overlap between gene enrichment categories. Although these measures are useful in their respective contexts, they cannot be generalized to other networks. In contrast, by incorporating cluster similarity into our model, we have a standard way of calculating these measures for any type of network.

3.3.5 MM algorithm and R software implementation

Our software implementation of CPBA is freely available in the R package PropClust. On a laptop with a 2.4 GHz i5 processor and 4 GB of RAM, PropClust can estimate the parameters for 1000 nodes for a given cluster assignment in 0.1 seconds. For 3000 nodes, the same analysis takes 1 second. In practice, initial clusters are never perfect and must be re-configured as well. PropClust adopts a block descent (or ascent) strategy that alternates cluster re-assignment and parameter re-estimation until clusters stabilize. Block descent takes under

10 rounds on average if initial cluster assignments are good. Note that all parameters are fixed in cluster re-assignment, and all clusters are fixed in parameter re-estimation. Furthermore, both steps decrease the value of the objective function. Early versions of PropClust re-estimated parameters as each node was moved. This tactic proved to be too computationally burdensome on large-scale problems despite its slightly better performance in finding optimal clusters.

Judicious choice of the initial clusters is realized by a divide-and-conquer strategy. First, hierarchical clustering coupled with dynamic branch cutting [LZH07] is used to cluster nodes into manageable blocks of user-defined maximum size, for instance at most 1000 nodes each. Second, the CPBA algorithm is applied to each block to arrive at clusters within blocks. Our co-expression network application shows that this initialization procedure works well even in large data sets. Another way to accelerate clustering is exploit parallel processing in the MM algorithm. Parallelization of the MM algorithm is easily achieved since the parameters separate in the surrogate function and updating the propensities via (Eq. 3.12) and (Eq. 3.15) requires only the previous parameter values. Cluster re-assignment avoids continuous optimization altogether and is very fast.

3.3.6 Simulated clusters in the Euclidean plane

Our first simulated dataset suggests a geometric interpretation of propensities and cluster similarities. For this dataset we simulated four distinct clusters on the Euclidean plane by sampling from a rotationally symmetric normal distribution with covariance matrix $\boldsymbol{I}$ and means corresponding to the four cluster centers shown in Figure 3.1a. The numbers of points in the four clusters were 50, 100, 150, and 200, respectively. The adjacency between two points is defined as $1 - [\text{dist}/\max(\text{dist})]^2$, where dist denotes Euclidean distance between the points and max(dist) denotes the maximum distance between any two points. Thus as depicted in Figure 3.1b, points closer together have a higher adjacency than those further apart. As anticipated, the MM algorithm provided the correct cluster assignments. Figure 3.1c also makes it evident that the estimated propensity of a point is significantly correlated

to the Euclidian distance between the point and its cluster's center. This result is expected since a connectivity k_i is related to a propensity p_i through equation (Eq. 3.7). Within a module, connectivity is also related to its cluster's center through the formula

$$k_i \;=\; (n-1) - \frac{n||\boldsymbol{x}_i - \bar{\boldsymbol{x}}||^2 + \sum_j ||\boldsymbol{x}_j - \bar{\boldsymbol{x}}||^2}{\max(\text{dist})}$$

where n is the number of nodes in the cluster, $\boldsymbol{x}_i$ is the position of node i, and $\bar{\boldsymbol{x}}$ is the position of the cluster center [Hor11]. This formula also explains why there is a separate line for each cluster in Figure 3.1c. Finally, Figure 3.1d shows that the cluster similarity r_{kl} of clusters k and l is significantly correlated to the distance between the centers of k and l. In summary, a propensity can be viewed as a measure of the centrality of a node, while a cluster similarity reflects the distance between two cluster centers.

3.3.7 Simulated gene co-expression network

To illustrate how CPBA generalizes to weighted correlation networks, we simulated gene expression data using the `simulateDatExpr5Modules` function in the WGCNA package in R [LH08]. Given the simulated expression data, we calculated Pearson's correlation coefficient for each pair of genes and formed an adjacency matrix. Applying CPBA based clustering to the simulated data led to clusters that overlap perfectly with the simulated clusters. As Figure 3.2 depicts, the estimated propensities p_i are very significantly correlated to the node connectivities k_i. This strong relationship reflects (Eq. 3.7). Furthermore, as seen in Figure 3.2, cluster similarity is significantly correlated to true eigengene adjacency, namely, $r_{c_i c_j} \approx |\operatorname{Corr}(\boldsymbol{u}_1^{c_i}, \boldsymbol{u}_1^{c_j})|^\beta$. In both simulations several different cluster assignment initializations were tried and all led to the same, correct, result.

3.3.8 Real gene co-expression network application to brain data

In this real data example, we demonstrate that CPBA generalizes weighted correlation network analysis and can deal with fairly large data sets. The human brain expression data in question were measured on the Affymetrix U133A platform [OKI08]. Following Oldham et al 2008 [OKI08], we restrict our analysis to the roughly 10^4 probes that were highly ex-

pressed in brain tissue. The biological modules discovered by Oldham et al 2008 [OKI08] via WGCNA are fairly well understood and correspond to cell types such as astrocytes, oligodendrocytes, and neurons enriched for specific cell markers. In re-analyzing these data, we defined initial clusters as sketched in our discussion of the R software implementation of CPBA. This strategy obviates the need to pre-specify the number of clusters present in a data set. The results of PropClust are depicted in the second color band of Figure 3.3A. Overall, we find that CPBA yields modules very similar to those identified by WGCNA. The overlap with the well annotated modules of Oldham et al [OKI08] shows that the two clustering procedures yield meaningful and nearly equivlaent modules. CPBA has the advantage of giving cluster similarities. Figure 3.3B shows that eigengene based network adjacency (defined as the correlation between eigengenes raised to the soft-thresholding power 4) is highly correlated (r=0.93) with the cluster similarity parameter calculated by CPBA. For genes within a given module, Figures 3.3C-E demonstrate that the node propensities estimated under CPBA are highly correlated with the module membership measures kME raised to the soft thresholding power 4. Finally, Figures 3.3I-J show that the connectivities k_i in the correlation network are highly correlated (r=0.96) with the connectivities calculated under the CPBA approximation and with the corresponding CPBA propensities (r=0.88).

These results demonstrate that CPBA is roughly equivalent to WGCNA in a typical co-expression network. We expect that CPBA will also be helpful in understanding network topology. For example, Figure 3.3F shows that the weighted co-expression network satisfies the approximate scale-free topology (SFT) property defined in (Eq. 3.10). Future research should aim to characterize the general fit of CPBA parameters to the SFT property. In this example, the CPBA based connectivities and propensities shown in Figures 3.3G and 3.3H agree well.

3.3.9 OMIM disease and gene networks

Here we present an application that is not amenable to correlation network models but is arguably well suited for multigraph models. Specifically, we consider a bipartite multigraph

between genes and diseases based on curated data from the reference Online Mendelian Inheritance of Man (OMIM), which tracks published links between diseases and corresponding genes [Gen]. These data were previously studied in detail by Goh et al. [GCV07], who showed that diseases and their associated genes are related at higher levels of cellular and organ function. In the current application we validated their functional clustering using the CPBA model.

Following Goh et al. [GCV07], we analyzed the data in two ways. First we created a disease network by placing an edge between two diseases for each gene they were both linked to. Only the links labeled as high quality by OMIM were considered. This construction yielded a multigraph of 2552 diseases with 1401 diseases connected to at least one other disease. We created a second, complementary multigraph by placing an edge between two genes for each disease they were both linked to. For this multigraph, there were 4045 genes with 1978 genes connected to at least one other gene. As suggested by the Medical Subject Headings (MeSH) list [Rog63], we applied the CPBA model with $K = 10$ clusters for the gene network and $K = 14$ clusters for the disease network, leaving out irrelevant categories.

We categorized the diseases using MeSH with little success. Nearly half of the diseases (47%) were not mapped to any category, and another 36% were mapped to multiple categories. Using the clustering obtained from the CPBA analysis of the disease network, we looked at whether any MeSH categories were overrepresented in a cluster. Ignoring diseases present in multiple MeSH categories, we found 8 significant categories at $P < 0.01$, including neoplasms, musculoskeletal diseases, and eye diseases (See Table 1). Although significant results were obtained, only a handful of diseases in each cluster contributed to the statistic. Upon closer inspection of the clusters, we found that many seemingly well-defined diseases were either not mapped or multiply mapped. For example, the eye disease cluster contains morning glory disc anomaly, coloboma, best macular dystrophy, cone-rod retinal dystrophy, and iris hypoplasia which are all clearly eye diseases, but not classified as such by MeSH. The eye disease cluster is depicted in Figure 3.4.

Additionally, we found 540 significant connections between diseases at $P < 0.01$ and 148 significant connections at $P < 0.001$. The top 10 connections are listed in Table 2. The

Table 3.1: Over-represented MeSH categories in the disease network.

Name	**MeSH Num.**	**-Log$_{10}$(P)**
Hemic & Lymphatic diseases	C15	8.32
Eye diseases	C11	7.78
Cardiovascular diseases	C14	4.23
Nervous System Diseases	C10	4.04
Neoplasms	C4	3.37
Musculoskeletal diseases	C5	2.91
Endocrine system diseases	C19	2.04
Congenital, Hereditary, & Neonatal diseases & Abnormalities	C16	2.03

disease pair Adrenoleukodystrophy and Zellweger syndrome came in first; these two diseases are two of only three peroxisome biogenesis diseases belonging to the Zellweger spectrum [SDR06]. It is also interesting to look for highly connected hub clusters, namely, clusters with high similarity to several other clusters. To define a measure of **cluster connectivity**, one can use the row sum of the cluster similarity matrix $\boldsymbol{R}$. The neoplasm cluster has the highest row sum and is the cluster with the highest cluster connectivity. This makes sense given the complexity and diversity of cancers within the cluster.

Looking at the complementary gene network, we checked for overrepresentation of Gene Ontology (GO) terms using BinGO on Cytoscape [MHK05]. We found that each cluster had an overrepresentation of many GO terms. In the cluster with the well-known tumor suppressor protein TP53, we found 875 statistically significant GO terms at $P < 0.01$. Of these, 585 terms are still significant at $P < 0.001$ after accounting for multiple testing. The top 10 GO terms include both positive and negative regulation of cellular and biological processes, regulation of cell proliferation, anatomical structure development, regulation of apoptosis, and others that are clearly associated with TP53. Finally, we found 1316 significant connections between genes at $P < 0.01$ and 418 at $P < 0.001$. The top 10 connections are listed in Table 3. Many of these gene pairs are known to interact from other supporting evidence. For example, interaction between the top ranking pair, Hemoglobin Alpha 1 globin chain

Table 3.2: Disease network top 15 significant connections CPBA.

	DISEASE 1	**DISEASE 2**	**C1**	**C2**	**-Log$_{10}$(P)**
1	Zellweger syndrome	Adrenoleukodystrophy	14	14	8.57
2	Muscular dystrophy-dystroglycanopathy (limb-girdle)	Muscular dystrophy-dystroglycanopathy (congenital)	2	2	7.05
3	Ullrich congenital muscular dystrophy	Bethlem myopathy	14	14	6.48
4	Iminoglycinuria	Hyperglycinuria	14	14	6.48
5	Alport syndrome	Hematuria	14	14	5.31
6	Colorblindness	Blue cone monochromacy	14	14	5.31
7	Refsum disease	Zellweger syndrome	14	14	5.05
8	Usher syndrome	Retinitis pigmentosa	8	6	5.04
9	Seckel syndrome	Microcephaly	14	14	4.96
10	Leukoencephalopathy with vanishing white matter	Ovarioleukodystrophy	14	14	4.96
11	Omenn syndrome	Severe combined immunodeficiency	14	14	4.90
12	Tuberous sclerosis	Lymphangioleiomyomatosis	14	14	4.60
13	Cone-rod dystrophy	Macular degeneration	6	10	4.60
14	Bronchiectasis with or without elevated sweat chloride	Pseudohypoaldosteronism	11	11	4.47
15	Leri-Weill dyschondrosteosis	Langer mesomelic dysplasia	14	14	4.10

(HBA1) and Hemoglobin Subunit Beta (HBB), is confirmed by their co-crystal structure in x-ray crystallography [Sha83] and by automated yeast two-hybrid (Y2H) interaction mating [Ste05]. Figure 3.5 depicts the full gene network derived from OMIM.

Table 3.3: Gene network top 20 significant connections CPBA.

RANK	**GENE 1**	**GENE 2**	**CLUSTER 1**	**CLUSTER 2**	**-Log$_{10}$(P)**
1	HBB	HBA1	2	2	9.05
2	SHOXY	SHOX	10	10	7.36
3	BDNF	HTR2A	5	4	7.07
4	SH2B3	JAK2	2	8	7.05
5	TSC2	TSC1	10	10	6.28
6	FOXC1	PITX2	7	7	5.73
7	MAPT	PSEN1	4	6	5.66
8	OPN1MW	OPN1LW	10	10	5.58
9	COL4A4	COL4A3	10	10	5.58
10	RAG2	RAG1	10	10	5.56
11	SCNN1G	SCNN1B	5	5	5.25
12	HBB	KLF1	2	10	5.09
13	COL6A1	COL6A3	10	10	5.08
14	COL6A2	COL6A3	10	10	5.08
15	SLC6A19	SLC36A2	10	10	5.08
16	SLC6A20	SLC36A2	10	10	5.08
17	SLC6A20	SLC6A19	10	10	5.08
18	COL6A2	COL6A1	10	10	5.08
19	GPC3	OFD1	8	7	4.75
20	LTBP2	CYP1B1	10	7	4.73

3.3.10 Empirical comparison of edge statistics

In this section we compare our current CPBA model with our original Pure Propensity Poisson (PPP) model on two real datasets: the OMIM disease network and the complimentary OMIM gene network. On the whole we find that the CPBA model produces more plausible P-values for the edge-count tests. Conditioning on clusters enables CPBA to detect significant intercluster connections often missed by the PPP model. It also produces more reasonable P-values within clusters since propensities are not artificially deflated by the lack of connections between nodes from different clusters. We now consider how these trends play out in the OMIM disease network and the OMIM gene network.

In the disease network we find that, among the 20 most significant connections under the CPBA model, 5 are intercluster connections (See Table 2). Under the PPP model in contrast, none of the 20 most significant connections link different CPBA clusters (See Table 5). In fact, none of the top 50 connections of the PPP model occur between different CPBA clusters. The significant connection between Usher syndrome and retinitis pigmentosa would have gone completely unnoticed under the PPP model. This would be a shame because retinitis pigmentosa is a major symptom of Usher syndrome [NIHb]. Another missed intercluster pairing, Waardenburg syndrome and Craniofacial-deafness-hand syndrome, also deserves recognition since both syndomes involve deafness and common facial features [NIHc, NIHa].

Comparing the intracluster connections, we find that CPBA and PPP produce similar results, with 8 connections present in both lists. However, the P-values of these connections differ sharply under the two models. Since the PPP model essentially assumes a single cluster, estimated propensities trend lower in response to the lack of connections between nodes from different clusters. This results in lower means for the Poisson distributions and more extreme P-values. This phenomenon is especially evident in the pairing between Adrenoleukodystrophy and Zellweger syndrome; in the CPBA model the test for excess edges has $-\text{Log}_{10}(P) = 8.57$, whereas in the PPP model $-\text{Log}_{10}(P) = 12.06$.

The same story holds for the gene network. Among the 20 most significant connections under CPBA, 7 are intercluster connections. Under the PPP model the corresponding num-

ber is 0 (See Tables 3 and 5). One of the more interesting missed connections occurs between BDNF (brain-derived neurotrophic factor) and HTR2A (seratonin receptor 2A). Both genes are associated with attention in schizophrenia [ALG08]. As for intracluster connections, all intracluster connections found in the CPBA list are also found in the PPP list. However, the P-values for the most significant pair (HBB and HBA) differ by almost 5 orders of magnitude.

To summarize, the CPBA model was able to find significant intercluster edge counts that the PPP model missed. Indeed, the PPP model was unable to find a single signficant intercluster pair in either data set. Although conditioning on clusters resulted in less impressive intracluster P-values, the CPBA model was still able to detect most of the significant intracluster pairings found by the PPP model. Figure 3.6 provides a scatterplot of $-\text{Log}_{10}(P)$ for all significant connections obtained under either model. Points are colored red if they represent a pairing within a cluster and black if they represent a pairing between different clusters. The figure justifies our contentions that the CPBA model is more sensitive to intercluster connections and less sensitive to intracluster connections than its less nuanced competitor. So while there will be fewer significant intracluster connections, they will arguably be more interesting. Most likely these virtues of the CPBA model carry over to other data sets.

3.3.11 Simulations for evaluating edge statistics

To drive home the last point, we took a block diagonal adjacency matrix containing 1's in its diagonal blocks and 0's in its off-diagonal blocks and introduced a few off-block connections. In our initial matrix with three diagonal blocks of 100, 200, and 500 nodes, we changed 60 off-block entries from 0's to 1's. Each pair of node sets accounted for 20 of these switches. We then analyzed the modified matrix under both the CPBA and PPP models. Figure 3.7 plots $-\text{Log}_{10}(P)$ versus true adjacencies for the modified entries. Based on its identification of clusters, the CPBA model yields a better fit to the data. Comparison of $-\text{Log}_{10}(P)$ values under the two models shows that CPBA is more adept at finding significant intercluster connections. The evidence from the receiver operating characteristic (ROC) curve is very convincing on this point. The area under the ROC curve for the CPBA model was 0.95

compared to just 0.38 for the PPP model.

3.3.12 Hidden relationships between Fortune 500 companies

To illustrate the utility of CPBA in a non-biological setting, we briefly describe a multigraph model of cross-company management. Specifically, we took the Fortune 500 Companies of 2011 and put an edge between two companies for each shared member on their boards of directors. The original data is found in Freebase [fre]. As discussed below, the use of the Bayesian Information criterion (BIC) or the Akaike Information criterion for estimating clusters is problematic. For example, the BIC suggests an optimal number of clusters K around 10, while the AIC gives a less plausible value of $K > 20$. In the following, we assume $K = 10$ clusters. It is noteworthy that most companies do not cluster into groups of related industries. This makes sense because conflict of interest norms preclude companies in the same field from sharing board members. Overt clustering is consequently discouraged.

Based on the underlying probability model, we ascertained the significance of the edge counts for company pairs. Table 5 lists the 10 most significant connections under the 10-cluster model. Several connections stand out. The significant pairing between Fidelity National Financial and Fidelity National Information Services is rather obvious. The same holds for the pairing between Autozone and AutoNation Inc. Other connections are less obvious. The pairing between General Motors and DuPont may reflect the fact that Pierre du Pont, the founder of DuPont, at one point owned a third of all General Motors stock. This remained true until federal antitrust prosecutors filed suit, and the Supreme Court ruled against DuPont, forcing the company to dispose all of its GM shares in 1961 [dup]. Although the shares are gone, it seems that some ties persist.

3.3.13 Relationship to other network models and future research

Because so much effort has been devoted to the mathematical and statistical explication of complex networks, we can only touch on the relationship of the CPBA parametrization to other network models. Complex networks can be described by random graphs (the Erdös

and Rényi model [ER59]), small-world models (the Watts Strogatz model [WS98]), scale-free networks (the preferential attachment model of Barabasi and Albert [BA99, AB02]), and other growing random network (GRN) models. These models involve graphs rather than multigraphs, so the number of edges per node pair equals 0 or 1. The CPBA has interesting ramifications for random graphs with arbitrary connectivity distributions [NSW01]. If the edges are placed randomly in a network with such connectivities, then the probability P_{ij} of observing an edge between nodes i and j is exactly factorizable. In fact, $P_{ij} = k_i^{-3} k_j^{-3}$, where k_i is the connectivity (degree) of node i [KRL00, AB02]. Thus, P_{ij} can be well approximated by CPBA with propensities $p_i = k_i^{-3}$ and cluster similarities $r_{ab} = 1$. The Erdös and Rényi (ER) model, which assumes uniform edge probabilities, is too restrictive for realistic networks. The CPBA parametrization adapts well to random graphs if we replace the mean edge count with the edge formation probabilities

$$P_{ij} \quad = \quad \frac{p_i p_j r_{c_i c_j}}{1 + p_i p_j r_{c_i c_j}}.$$

This reformulation of the model is consistent with construction of an MM algorithm for parameter estimation [Lan10]. Future research should explore the topological properties of such models.

Growing random networks (GRNs) are also of interest since many networks grow by the continuous addition of new nodes and exhibit preferential attachment. Thus, the likelihood of connecting to a node depends on the node's current connectivity [BA99, KRL00, NSW01, Str01, AB02, Dur06]. At each time step of a growing random network [KRL00], a new node is added, and a directed edge to one of the earlier nodes is created. This growing network has a directed-tree graph topology whose basic elements are nodes connected by directed edges. In general, the topology of a general GRN is determined by the connection kernel A_k, which is the probability that a newly-introduced node forms an edge to an existing site with k edges ($k-1$ incoming and 1 outgoing). Future research could explore how to define a connection kernel (or more generally a GRN) so that the resulting network can be well approximated using the CPBA of the adjacency matrix. The Barabasi-Albert (BA) model is an important special case of a GRN [BA99, AB02] that leads to a scale-free network. In

the BA model, the degree of a node satisfies the power-law (or scale-free) distribution

$$P(k) \sim k^{\gamma} \,. \tag{3.10}$$

For homogeneous connection kernels, $A_k \sim k^{\nu}$, and scale free networks only arise if $\nu = 1$ [KRL00]. Future research could explore whether the adjacency matrix of the BA model can be well approximated using the CPBA. Toward this end it may useful to observe that the probability P_{ij} of finding an edge between nodes i and j in the BA model is given by [KRL00, AB02]

$$P_{ij} \;=\; \frac{4(k_j-1)(4k_i+k_j+2)}{k_i(k_i+1)(k_i+k_j-1)(k_i+k_j)(k_i+k_j+1)(k_i+k_j+2)}$$

which, importantly, assumes that node i with connectivity k_i was added later to the growing network than node j (implying that $k_i < k_j$). In view of this temporal assumption, P_{ij} is not symmetrical in i and j; it also contains no parameters to capture clustering. Thus, there is no obvious relationship between the BA model and the CPBA approximation of a network. Future research can investigate how to parameterize preferential attachment so that the resulting probability P_{ij} of finding an edge fits well to the CPBA.

3.3.13.1 Relationship to other clustering methods

Although the MM algorithm that estimates the CPBA parameters naturally generates a clustering method, CPBA is not just another clustering method. Our applications highlight the utility of the parameter estimates and the resulting likelihood based tests. CPBA not only provides a sparse parametrization of a general similarity matrix, but it also identifies hub nodes and clusters and enables significance tests for excess edges (between nodes) and shared similarities (between clusters). We do not claim that CPBA based clustering outperforms existing clustering methods in the simple task of clustering.

Substitutes for include CPBA clustering include hierarchical clustering, partitioning around medoids [KR90], spectral clustering [ZB07], mixture models [NL07], component models [SAK08], and many more [HW08, KTG06, ABF08, New06, Sch07]. Because CPBA can be interpreted as a generalization of weighted correlation network methods, there is no need

to invoke it instead of WGCNA when it comes to co-expression network applications. In modeling relationships between quantitative variables, one can use a host of other methods, for example sparse Gaussian graphical models [YL11, XL10], Bayesian networks, and structural equation models. CPBA is not meant to replace these powerful approaches for modeling relationships between quantitative variables. Its main attraction is that it applies to a general similarity measure. Since input data sometimes consists of a similarity (or dissimilarity) measure, CPBA fills a useful niche.

3.4 Conclusions

The current paper introduces the CPBA model (cluster and propensity based approximation) for general similarity measures and sketches an efficient MM algorithm for estimation of the CPBA parameters. These advances will prove valuable in dissecting networks involving functional or evolutionary modules. The CPBA model is attractive for several reasons. First, it invokes relatively few parameters while providing sufficient flexibility for modeling observed similarities. Second, the cluster similarity parameters are good at revealing higher-order relationships between clusters. The row sum of the cluster similarity matrix can be used to define a cluster connectivity measure and to identify hub clusters such as the neoplasm hub in the disease network. Third, the CPBA model naturally generalizes network approximations that are already part of scientific practice, namely, the propensity based approach in multigraph models, the conformity decomposition in weighted networks, and the eigenvector based approximation in correlation networks. Fourth, the connections to the MM algorithm make the model well adapted to high-dimensional optimization. Fifth, the Poisson multigraph version of the model enables assessment of the statistical significance of edge counts and similarities between clusters. Sixth, likelihood-based models such as the Poisson multigraph model provide a rational basis for estimating the number of clusters. While it is beyond our scope to evaluate different methods for estimating the number of clusters in a data set, it is worth mentioning that our R implementation allows users to initialize clusters via hierarchical clustering. This tactic obviates the need to pre-specify the

number of clusters.

Using simulated clusters in the plane and simulated co-expression networks, we demonstrate that CPBA generalizes existing methods. The planar examples show how a propensity can be intuitively seen as a measure of a node's closeness to its cluster's center and how a cluster similarity can be seen as a measure of proximity between two clusters. The simulated gene expression dataset exposes the CPBA model's close ties to the previously studied concepts of intramodular connectivity, module eigengenes, and eigengene adjacency. Our analysis of real gene expression data reassures us that CPBA clustering results are similar to those of a benchmark method used in co-expression network analysis. The CPBA propensity parameters mirror the module eigengene based connectivity kME, and the cluster similarity measures mimic the network eigengenes. In our view, the main value of the CPBA model lies in generalizing correlation network methods.

To illustrate the versatility of CPBA, we applied it to the gene and disease networks of OMIM. The evidence that CPBA identifies biologically meaningful clusters is readily apparent in the significant enrichment of MeSH categories in the disease clusters and in the significant enrichment of GO terms in the gene clusters. While many other clustering procedures could have been used, CPBA has the advantage of dealing with dissimilarity measures as opposed to numeric input variables. It also provides Poisson likelihood based significance tests for edge counts (either pairs of genes and or pairs of diseases) that respect the underlying cluster structure. Finally, the row sums of the cluster similarity measure can be used to define hub clusters, and the estimated propensities can be used to define hub nodes. As we hoped, there were biologically meaningful ties between significantly connected pairs of genes and diseases. Several of these biologically plausible explanations are discussed in the text.

Although our examples are mainly biological, one can apply CPBA in many other contexts. For example, we employed CPBA to highlight shared board members among the Fortune 500 companies. This example illustrates how significant connections mirror the underlying ties between nodes. The edge count significance test suggests that the antitrust suit against GM and DuPont was no accident. To its credit, CPBA not only generalizes correla-

tion network methods to general similarity matrices, but it also provides a valuable extension of random multigraph methods to weighted and unweighted multigraph data. CPBA is not just another clustering procedure but offers unique test statistics that permit identification of hub clusters and significant edge counts. We anticipate that the CPBA model will prove attractive to a wide range of scientists.

3.5 Methods

3.5.1 Maximizing the Poisson log-likelihood based objective function

Our algorithm for maximizing the Poisson log-likelihood (Eq. 3.6) given a clustering assignment $\boldsymbol{c}$ combines block ascent and the MM principle [HL04, Lan04, WL10]. Clustering proceeds by re-assigning each node in turn until clusters stabilize. It may take several cycles through the nodes to achieve stability. Reassignment fixes parameters and selects the assignment with the highest log-likelihood. In the Poisson log-likelihood (Eq. 3.6), we take

$$\ln L(\boldsymbol{c}, \boldsymbol{R}, \boldsymbol{P}) \;=\; \sum_i \sum_{j \neq i} n_{ij} \left[\ln(r_{c_i c_j} p_i p_j) - (r_{c_i c_j} p_i p_j) - \ln(n_{ij}!) \right]$$

where $r_{c_i c_j}$ is the cluster similarity between clusters c_i and c_j, p_i is the propensity of node i, and $A_{ij} = n_{ij}$ is the number of connections between nodes i and j.

To optimize the objective function for a given cluster assignment, we employ block ascent and alternate updating $\boldsymbol{R}$ and $\boldsymbol{p}$. Fixing $\boldsymbol{p}$, it is possible to to solve for the best cluster similarity parameters r_{ab} exactly. Indeed, setting the partial derivative

$$\frac{\partial}{\partial r_{c_i c_j}} \text{Poisson} \;=\; \sum_{k \in c_i} \sum_{l \in c_j} \left(\frac{n_{kl}}{r_{c_i c_j}} - p_k p_l \right).$$

equal to zero and solving for $r_{c_i c_j}$ yields the simple update

$$\hat{r}_{c_i c_j} \;=\; \frac{\sum_{k \in c_i} \sum_{l \in c_j} n_{kl}}{(\sum_{k \in c_i} p_k)(\sum_{l \in c_j} p_l)}. \tag{3.11}$$

We expect the estimated r_{ab} to occur within the unit interval $[0, 1]$ because edge formation is enhanced within clusters.

To update the propensity vector $\boldsymbol{p}$ with $\boldsymbol{R}$ fixed, we turn to an MM algorithm [HL04, Lan04, WL10]. The MM principle says we should minorize the objective function by a surrogate function and maximize the surrogate function. This action drives the objective function uphill. One function minorizes another at a point $\boldsymbol{p}_m$ if it is tangent to the other function at $\boldsymbol{p}_m$ and falls below it elsewhere. The arithmetic-geometric mean inequality

$$p_i p_j \quad \leq \quad \frac{1}{2} p_{mi} p_{mj} \left[\left(\frac{p_i}{p_{mi}} \right)^2 + \left(\frac{p_j}{p_{mj}} \right)^2 \right]$$

is the key to minorizing the Poisson log-likelihood. Substituting the right-hand side for $p_i p_j$ in the log-likelihood (Eq. 3.6) gives a surrogate function with parameters separated and leads directly to the propensity updates

$$p_{m+1,i} \quad = \quad \sqrt{\frac{p_{mi} \sum_j n_{ij}}{\sum_{j \neq i} r_{c_i c_j} p_{mj}}} \,. \tag{3.12}$$

In practice, this MM algorithm may require an excessive number of iterations to converge. To accelerate convergence, we employ a Quasi-Newton extrapolation specifically designed for high-dimensional problems (Methods and [ZAL11]). The overall ascent algorithm (outer iterations) on $\boldsymbol{R}$ and $\boldsymbol{p}$ may also be slow to converge. It can also be accelerated by the same extrapolation scheme. Accelerating both inner and outer iterations gives a fast numerically stable procedure for estimating $\boldsymbol{R}$ and $\boldsymbol{p}$ for $\boldsymbol{c}$ fixed.

3.5.2 Minimizing the Frobenius norm based objective function

Minimization of the Frobenius objective function (Eq. 3.5) employs block descent and again alternates updating $\boldsymbol{R}$ and $\boldsymbol{p}$. In this case setting the partial derivative

$$\frac{\partial}{\partial r_{c_i c_j}} \text{Frobenius} \quad = \quad -2 \sum_{k \in c_i} \sum_{l \in c_j} (A_{kl} - r_{c_i c_j} p_k p_l) p_k p_l$$

equal to zero and solving for $r_{c_i c_j}$ yields the simple update

$$\hat{r}_{c_i c_j} \quad = \quad \frac{\sum_{k \in c_i} \sum_{l \in c_j} A_{kl} p_k p_l}{(\sum_{k \in c_i} p_k^2)(\sum_{l \in c_j} p_l^2)} \,. \tag{3.13}$$

To update $\boldsymbol{p}$ for $\boldsymbol{R}$ fixed, we again rely on the MM principle. However, since we now seek to minimize the objective function, we majorize it. This is accomplished by first expanding

it in the form

$$\text{Frobenius}(\boldsymbol{c}, \boldsymbol{p}, \boldsymbol{R}) \;=\; \sum_i \sum_{j \neq i} \left[A_{ij}^2 - 2A_{ij} r_{c_i c_j} p_i p_j + (r_{c_i c_j} p_i p_j)^2 \right] . \tag{3.14}$$

In majorization, one is allowed to work piecemeal. Thus, we majorize the term involving $(p_i p_j)^2$ by the earlier arithmetic-geometric mean inequality

$$p_i^2 p_j^2 \;\leq\; \frac{1}{2}(p_{mi})^2 (p_{mj})^2 \left[\left(\frac{p_i}{p_{mi}} \right)^4 + \left(\frac{p_j}{p_{mj}} \right)^4 \right]$$

taking into account squares. The term involving $-p_i p_j$ can be majorized by the inequality $x \geq 1 + \ln x$ for $x \geq 0$ in the form

$$-p_i p_j \;\leq\; -p_{mi} p_{mj} \left[1 + \ln \left(\frac{p_i p_j}{p_{mi} p_{mj}} \right) \right] .$$

Substituting upper bounds side for $(p_i p_j)^2$ and $-p_i p_j$ in the expanded objective function (Eq. 3.14) gives a surrogate function with parameters separated and leads directly to the propensity updates

$$p_{m+1,i} \;=\; \left[\frac{(p_{mi})^3 \sum_b \sum_{j \in b} r_{c_i c_j} A_{ij} p_{mj}}{\sum_b \sum_{j \in b} (r_{c_i c_j})^2 (p_{mj})^2} \right]^{1/4} . \tag{3.15}$$

As in the Poisson case, acceleration is advisable for both inner MM iterations and the outer block descent iterations. The same Quasi-Newton extrapolation [ZAL11] is pertinent and gives a fast numerically stable procedure for estimating $\boldsymbol{R}$ and $\boldsymbol{p}$ for $\boldsymbol{c}$ fixed.

3.5.3 Model Initialization

3.5.3.1 Initial cluster assignment

Many algorithms exist for creating initial cluster assignments [Sch07]. For most datasets these assignments only affect the time to convergence and not the converged solution. Our R software implements hierarchical clustering and does not require pre-specifying the number of clusters. More specifically, our software applies average linkage hierarchical clustering with dynamic branch cutting [LZH07]. Dissimilarities are set equal to 1 minus similarities.

3.5.3.2 Initial Propensities

One way to initialize propensities is to assume a single cluster and estimate propensities as suggested in our earlier work [RAS10]. An alternative in the Frobenius model is to initialize p_i by the sum of the connections of node i divided by the square root of the sum of all connections [Hor11],

$$p_i = \frac{\sum_{j \neq i} A_{ij}}{\sqrt{\sum_k \sum_{j \neq k} A_{kj}}}. \tag{3.16}$$

This initialization can be motivated by showing that the above equation holds if $r_{c_i c_j} = 1$ (equivalently, the network consists of a single cluster) and $\sum p_i \gg \sum p_i^2$. While the assumption of perfect cluster similarity is unrealistic, it leads to initial values that work well in practice. For the Poisson model the analog is

$$p_i = \frac{\sum_{j \neq i} n_{ij}}{\sqrt{\sum_k \sum_{j \neq k} n_{kj}}}. \tag{3.17}$$

3.5.3.3 Cluster similarity parameters

Because the block updates (Eq. 3.11) and (Eq. 3.13) for the cluster similarity parameters only depend on cluster assignment and propensities, it is natural to use those updates for initialization as well.

3.5.4 Clustering algorithm

1. Choose the objective function (Frobenius or Poisson).
2. Initialize the cluster assignment, for example, via hierarchical clustering.
3. Initialize the propensity vector $\boldsymbol{p}$ by (Eq. 3.16) or (Eq. 3.17) and the cluster similarity matrix $\boldsymbol{R}$ by (Eq. 3.11) or (Eq. 3.13).
4. Parameter Estimation: Given cluster assignments, re-estimate parameters through the updates (Eq. 3.11) and (Eq. 3.12) or (Eq. 3.13) and (Eq. 3.15). Declare convergence when the objective function changes by less than a threshold, say 10^{-5}.

5. Cluster Reassignment:

   (a) Randomly permute the nodes.

   (b) For each node taken in order, try all possible cluster reassignments for the node.

   (c) Assign the node to the cluster that leads to the biggest improvement in the objective function.

   (d) Repeat step 5 until no nodes are reassigned.

6. Repeat steps 4 and 5 until no nodes are reassigned.

7. (Optional) Repeat steps 1- 5 for other cluster numbers and use a cluster number estimation procedure for choosing the number of clusters.

3.5.5 Quasi-Newton Acceleration

In this section we briefly review a Quasi-Netwon acceleration method described more fully in [ZAL11]. Newton's method seeks a root of the equation $\mathbf{0} = \boldsymbol{x} - F(\boldsymbol{x})$, where $F(\boldsymbol{x})$ is a smooth map. For CPBA this is the algorithm map summarized by Equations (3.11) and (3.12) for Poisson updates or Equations (3.13) and (3.15) for Frobenius updates. Because the function $G(\boldsymbol{x}) = \boldsymbol{x} - F(\boldsymbol{x})$ has differential $dG(\boldsymbol{x}) = \boldsymbol{I} - dF(\boldsymbol{x})$, Newton's method iterates according to

$$\boldsymbol{x}_{n+1} \;=\; \boldsymbol{x}_n - dG(\boldsymbol{x}_n)^{-1} G(\boldsymbol{x}_n) \;=\; \boldsymbol{x}_n - [\boldsymbol{I} - dF(\boldsymbol{x}_n)]^{-1} G(\boldsymbol{x}_n)\,.$$

Quasi-Netwon acceleration approximates $dF(\boldsymbol{x}_n)$ by a low-rank matrix $\boldsymbol{M}$ and explicitly forms the inverse $(\boldsymbol{I} - \boldsymbol{M})^{-1}$.

Construction of $\boldsymbol{M}$ relies on secants. We can generate a secant by taking two iterates of the algorithm starting from the current iterate $\boldsymbol{x}_n$. If we are close to the optimal point $\boldsymbol{x}_\infty$, then we have the linear approximation

$$F \circ F(\boldsymbol{x}_n) - F(\boldsymbol{x}_n) \;\approx\; M[F(\boldsymbol{x}_n) - \boldsymbol{x}_n]\,,$$

where $\boldsymbol{M} = dF(\boldsymbol{x}_\infty)$. We abbreviate the secant requirement as $\boldsymbol{M}\boldsymbol{u} = \boldsymbol{v}$, where $\boldsymbol{u} = F(\boldsymbol{x}_n) - \boldsymbol{x}_n$ and $\boldsymbol{v} = F \circ F(\boldsymbol{x}_n) - F(\boldsymbol{x}_n)$. To improve the approximation of $\boldsymbol{M}$, one can use

several secant constraints $\boldsymbol{M}\boldsymbol{u}_i = \boldsymbol{v}_i$ for $i = 1, \ldots, q$. These are expressed in matrix form as $\boldsymbol{M}\boldsymbol{U} = \boldsymbol{V}$. For our purposes the value $q = 6$ works well.

Provided $\boldsymbol{U}$ has full column rank q, the minimum of the strictly convex function $||\boldsymbol{M}||_F^2$ subject to the constraints $\boldsymbol{M}\boldsymbol{U} = \boldsymbol{V}$ is attained at $\boldsymbol{M} = \boldsymbol{V}(\boldsymbol{U}^t\boldsymbol{U})^{-1}\boldsymbol{U}^t$ [ZAL11]. Fortunately, a variant of Sherman-Morrison formula [Lan99] implies that the matrix $\boldsymbol{I} - \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{V}(\boldsymbol{U}^t\boldsymbol{U})^{-1}\boldsymbol{U}^t$ has the explicit inverse

$$[\boldsymbol{I} - \boldsymbol{V}(\boldsymbol{U}^t\boldsymbol{U})^{-1}\boldsymbol{U}^t]^{-1} \;=\; \boldsymbol{I} + \boldsymbol{V}[\boldsymbol{U}^t\boldsymbol{U} - \boldsymbol{U}^t\boldsymbol{V}]^{-1}\boldsymbol{U}^t.$$

Thus, the quasi-Newton acceleration can be expressed as

$$\begin{aligned}
\boldsymbol{x}_{n+1} &= \boldsymbol{x}_n - [\boldsymbol{I} - \boldsymbol{V}(\boldsymbol{U}^t\boldsymbol{U})^{-1}\boldsymbol{U}^t]^{-1}[\boldsymbol{x}_n - F(\boldsymbol{x}_n)] \\
&= \boldsymbol{x}_n - [\boldsymbol{I} + \boldsymbol{V}(\boldsymbol{U}^t\boldsymbol{U} - \boldsymbol{U}^t\boldsymbol{V})^{-1}\boldsymbol{U}^t][\boldsymbol{x}_n - F\boldsymbol{x}_n)] \\
&= F(\boldsymbol{x}_n) - \boldsymbol{V}(\boldsymbol{U}^t\boldsymbol{U} - \boldsymbol{U}^t\boldsymbol{V})^{-1}\boldsymbol{U}^t[\boldsymbol{x}_n - F(\boldsymbol{x}_n)].
\end{aligned}$$

This update involves inversion of the small $q \times q$ matrix $\boldsymbol{U}^t\boldsymbol{U} - \boldsymbol{U}^t\boldsymbol{V}$; all other operations reduce to matrix times vector multiplications.

3.5.6 Estimating the number of clusters

Estimating the number of clusters is the Achilles heel of cluster analysis. While this topic is beyond our scope, it is worth mentioning that an advantage of model based approaches is that likelihood criteria can be brought to bear. Since adding clusters entails more parameters, it is tempting to use the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) to estimate the number of cluster in the Poissom model [Aka74, Sch78]. Both of these criteria balance the tradeoff between the number of parameters and the fit of the model. Specifically these methods choose the number of clusters K that minimize $AIC = -2\ln(L) + 2c$ or $BIC = -2\ln(L) + c\ln(n)$, respectively, where c is the number of parameters, L is the likelihood, and n is the sample size. We caution the reader that AIC and BIC may be inappropriate for the present task because both criteria invoke strong assumptions. For example, AIC is derived by assuming a regular model, for instance, a linear model with Gaussian noise. Hence, AIC may be inappropriate for models with latent variables such as

cluster labels. BIC may be inappropriate because our approach is frequentist rather than Bayesian. A review of the limitations and utility of these criteria can be found in [Wat09].

3.6 Other

3.6.1 Availability and requirements

Project name: PropClust R package
Project home page: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WG
Operating system(s): Platform independent
Programming language: R
Licence: GNU GPL 3

The propensity based clustering method propensityDecomposition is implemented in the R package PropClust. The package also contains the function CPBAdecomp for carrying out the propensity decomposition of a network.

3.6.2 List of abbreviations

BA - Barabasi Albert

CPBA - Cluster and Propensity Based Approximation

ER - Erdos Renyi

GO - Gene Ontology

GRN - Growing Random Network

kME - Connectivity based on the Module Eigenvector or Eigengene

MeSH - Medical Subject Headings

MM - Minorization maximization or majorization minimization

PPP - Pure Propensity Poisson

SFT - Scale Free Topology

SVD - Singular Value Decomposition

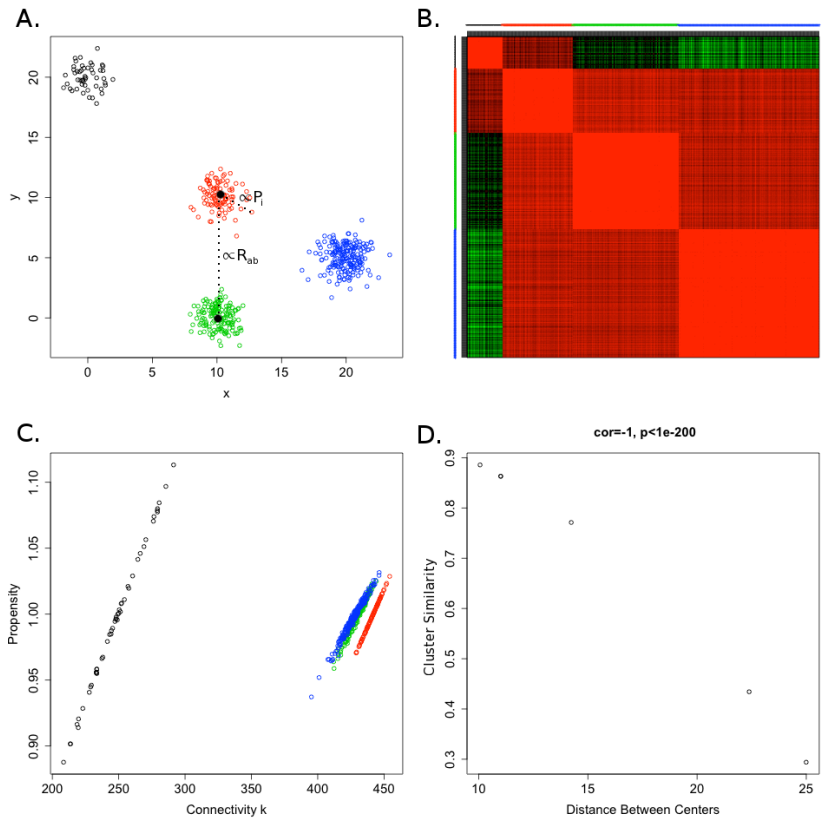


Figure 3.1: **Simulation providing a geometric interpretation of CPBA.** Four clusters were simulated in the Euclidean plane by sampling from the rotationally symmetric normal distribution with means corresponding to the different cluster centers and variance matrix I. The numbers of points in the clusters were 50, 100, 150, and 200 for the black, red, green, and blue clusters, respectively. A.) A plot of the points is shown colored by cluster. B.) Plots the resulting adjacency matrix, ordered by cluster, calculated using the formula 1[dist/max(dist)]2. In this plot red indicates a high adjacency, and green indicates a low adjacency. As expected, the adjacency within clusters is very high, and the adjacency between the blue and black clusters is the lowest since they are the furthest apart. C.) A plot of the relationship between propensity and connectivity as defined in (Eq. 7). These plots demonstrate geometrically how propensity is related to the distance between a point and its clusters center as expressed in (Eq. 10). D.) A plot that shows the relationship between cluster similarity and the distance between cluster centers. Note again the high correlation and significant p-value.

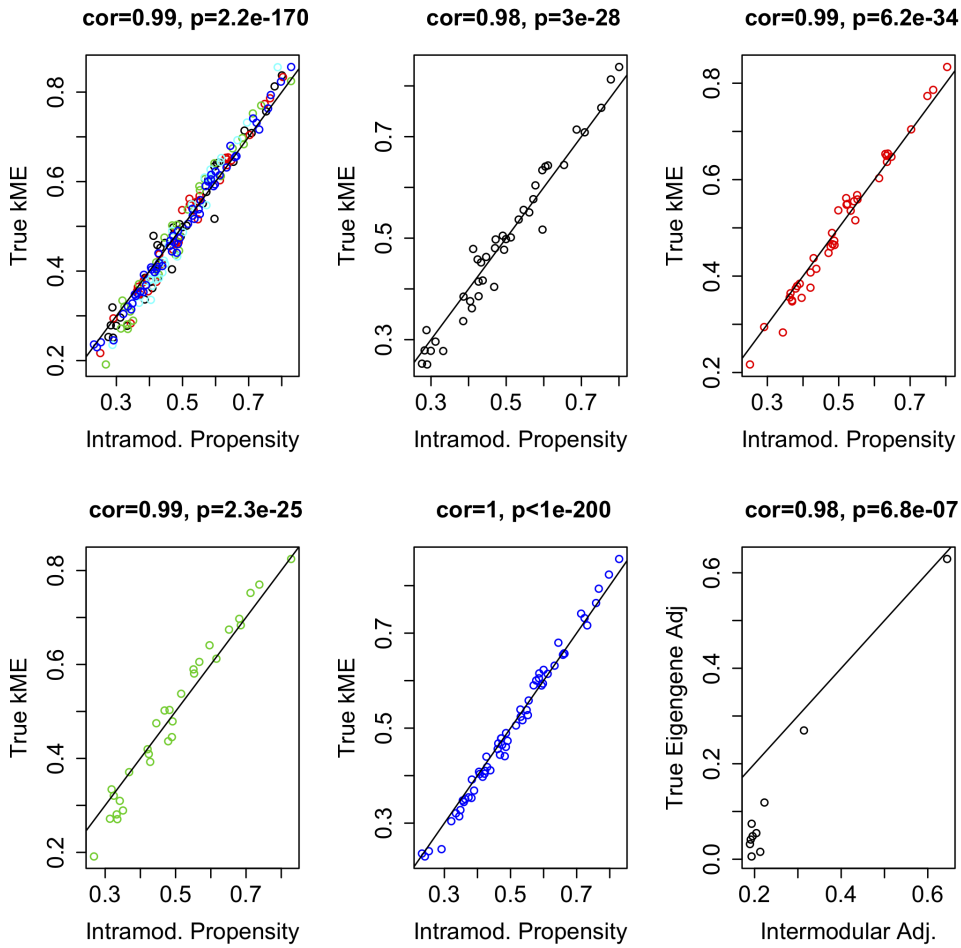


Figure 3.2: **Gene expression simulation results.** Gene expression data were simulated using the `simulateDatExpr5Modules` function under the WGCNA package in R. An adjacency matrix was then calculated from the Pearson correlation coefficients for the expression levels of each pair of genes. These plots reveal the relationship between the intramodular propensity and the true module membership, kME in (Eq. 3.3), first in all the clusters combined (top left) and then in each of the five clusters individually. Note the strong correlation and significant p-value in all cases.

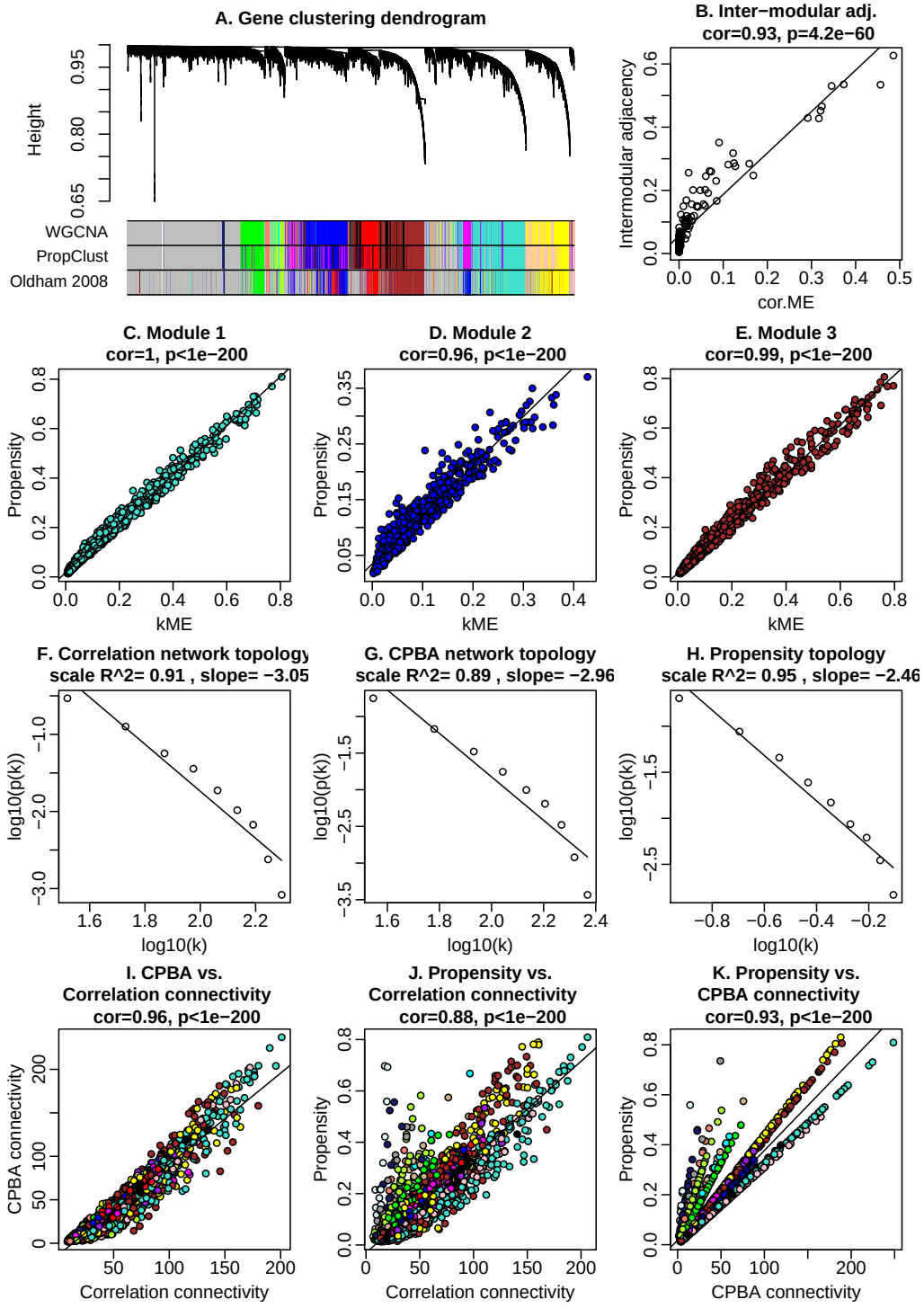


Figure 3.3: **Human brain expression data illustrate how CPBA can be interpreted as a generalization of WGCNA.** See section 3.3.8 for more information.

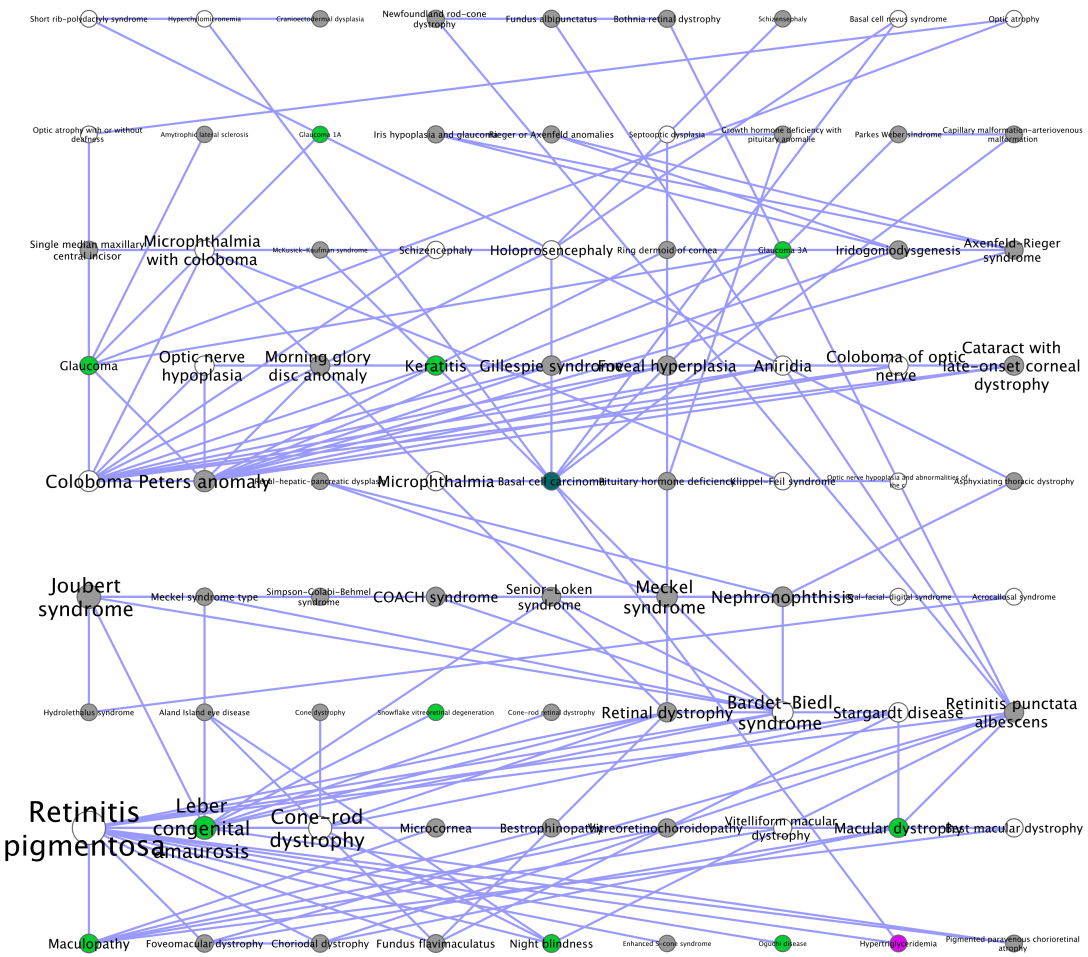


Figure 3.4: **OMIM disease network.** The intramodular connections between the nodes of the eye disease cluster are shown. Diseases are colored based on their MeSH categories, with diseases categorized as eye diseases (colored green), diseases linked to multiple categories (colored grey), and diseases that were not found (colored white). Note that more nodes should have been classified into the eye cluster by MeSH based on the name alone. Primary examples of this include retinitis pigmentosa, cone-rod dystrophy, retinal dystrophy, and microcornia. In spite of the failure of green labeling, these nodes were correctly classified by CPBA. Node and font sizes are proportional to a disease's propensity.

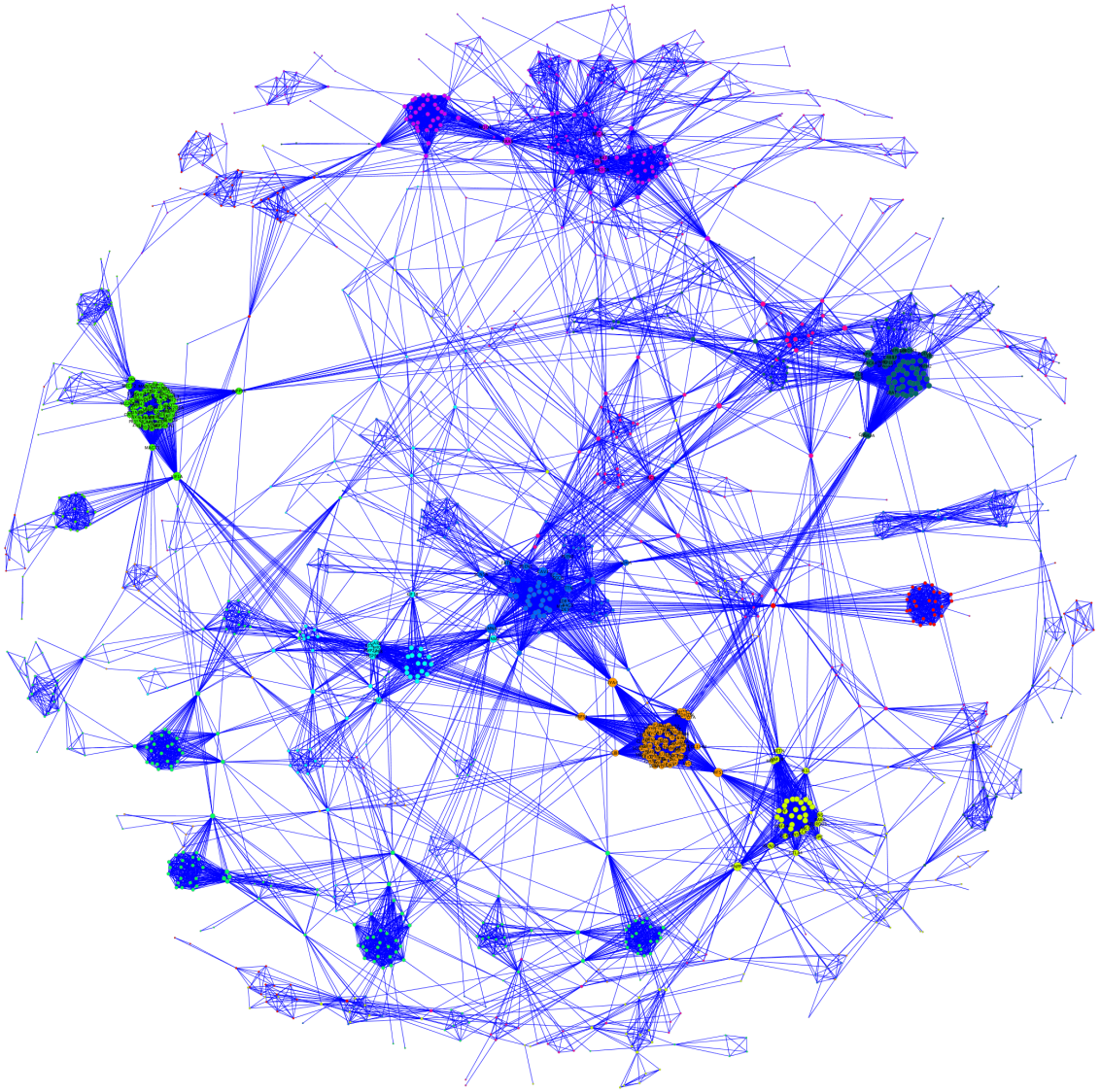


Figure 3.5: **OMIM Gene Network.** Genes are colored based on their cluster membership, and node size is proportional to a gene's propensity. This view was achieved with a spring-embedded layout in Cytoscape using the number of edges between two genes as weights. Note that CPBA based clustering identifies modules of highly interconnected nodes.

Table 3.4: Disease network top 15 significant connections PPP model.

	DISEASE1	**DISEASE2**	**C1**	**C2**	**-Log$_{10}$(P)**
1	Muscular dystrophy-dystroglycanopathy (limb-girdle)	Muscular dystrophy-dystroglycanopathy (congenital)	2	2	13.31
2	Zellweger syndrome	Adrenoleukodystrophy	14	14	12.06
3	Leber congenital amaurosis	Retinitis pigmentosa	6	6	10.12
4	Neuropathy	Charcot-Marie-Tooth disease	12	12	8.99
5	Blood group	Malaria	13	13	8.76
6	Ullrich congenital muscular dystrophy	Bethlem myopathy	14	14	8.57
7	Iminoglycinuria	Hyperglycinuria	14	14	8.57
8	Usher syndrome	Deafness	8	8	8.48
9	Hemolytic uremic syndrome	Macular degeneration	10	10	8.24
10	Bronchiectasis with or without elevated sweat chloride	Pseudohypoaldosteronism	11	11	7.75
11	Refsum disease	Zellweger syndrome	14	14	7.14
12	Meckel syndrome	Joubert syndrome	6	6	7.08
13	Omenn syndrome	Severe combined immunodeficiency	14	14	6.99
14	Left ventricular non-compaction	Cardiomyopathy	12	12	6.97
15	Mitochondrial complex I deficiency	Leigh syndrome	2	2	6.85

Table 3.5: Gene network top 20 significant connections PPP model.

RANK	**GENE 1**	**GENE 2**	**CLUSTER 1**	**CLUSTER 2**	**-Log$_{10}$(P)**
1	HBB	HBA1	2	2	13.87
2	SHOXY	SHOX	10	10	10.15
3	SDHD	SDHB	5	5	9.96
4	SCNN1G	SCNN1B	5	5	9.27
5	RAG2	RAG1	10	10	8.34
6	TSC2	TSC1	10	10	8.14
7	SDHC	SDHB	5	5	7.79
8	FOXC1	PITX2	7	7	7.54
9	OPN1MW	OPN1LW	10	10	7.43
10	COL4A4	COL4A3	10	10	7.43
11	GDF6	GDF3	7	7	7.29
12	TERC	TERT	9	9	7.20
13	CISH	TIRAP	4	4	7.12
14	GDNF	RET	5	5	7.04
15	COL6A1	COL6A3	10	10	6.94
16	COL6A2	COL6A3	10	10	6.94
17	SLC6A19	SLC36A2	10	10	6.94
18	SLC6A20	SLC36A2	10	10	6.94
19	SLC6A20	SLC6A19	10	10	6.94
20	COL6A2	COL6A1	10	10	6.94

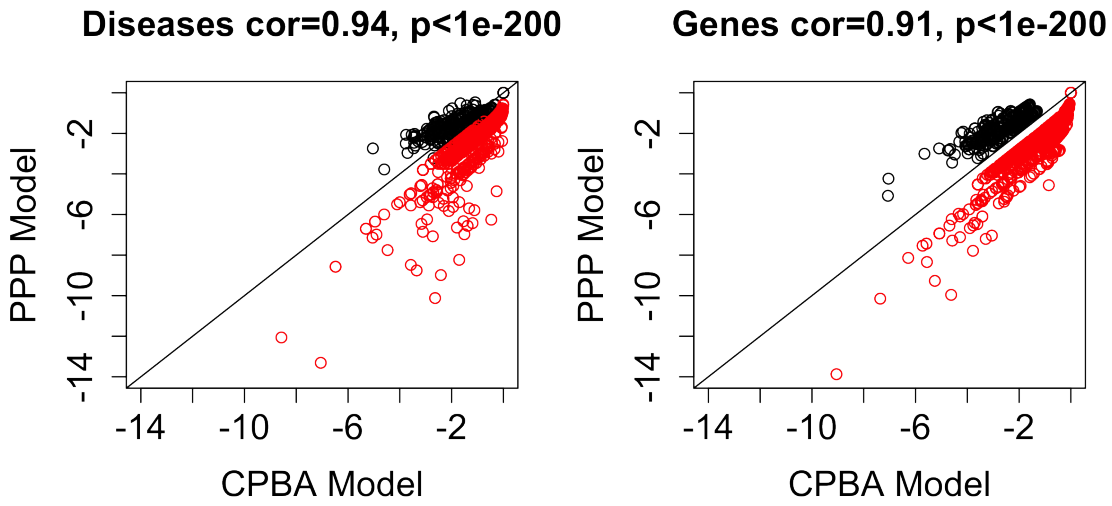


Figure 3.6: **OMIM CPBA versus PPP Analysis.** Scatterplot of the $\text{Log}_{10}(P)$ values obtained from analysis of OMIM using 14 and 10 clusters versus a single cluster for the Disease network and Gene network respectively. Note that the points are colored based on whether they come from a pair within a cluster(red) or between two clusters(black). This is very telling as it shows that by conditioning on the clustering, CPBA is able to increase its sensitivity in finding intercluster pairs while at the same time toning down that same trait in intracluster pairs.

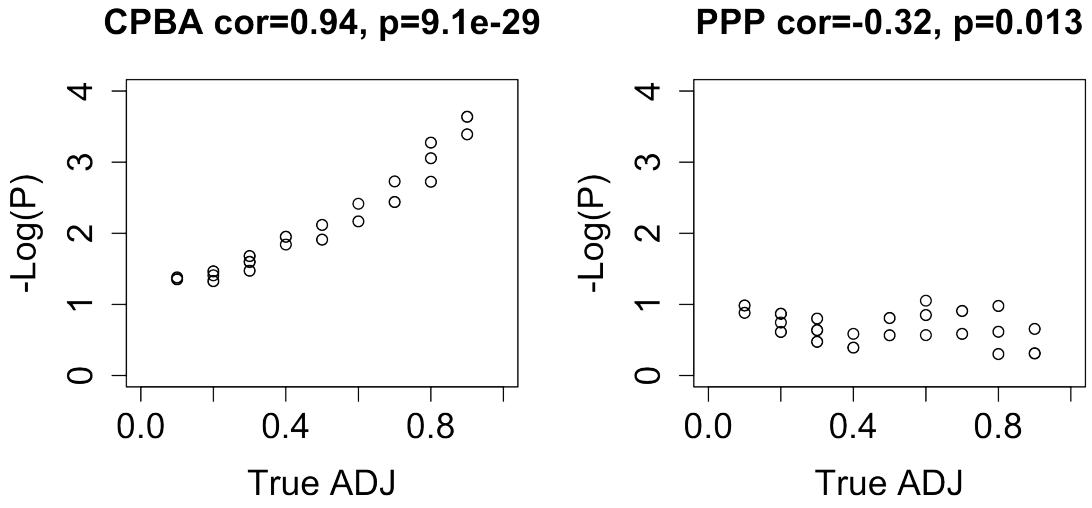


Figure 3.7: **Simulated CPBA versus PPP Analysis.** Scatterplot of the $-\text{Log}_{10}(P)$ values versus the true adjacency values obtained from 0/1 block diagonal matrix by re-setting a few other entries from 0 to 1. These changed values are shown along with the resulting $-\text{Log}_{10}(P)$ values obtained using CPBA and PPP.

Table 3.6: Fortune 500 top 10 significant connections.

RANK	**COMPANY 1**	**COMPANY 2**	**-Log$_{10}$(P)**	**Edges**
1	U.S. Bancorp	Ecolab	6.01	4
2	PetSmart	Dean Foods	4.53	3
3	Sempra Energy	Aecom Technology Corp.	4.39	3
4	General Motors	DuPont	4.07	3
5	Cardinal Health	Aon Corp.	4.07	3
6	Lockheed Martin	Monsanato	4.07	2
7	Fidelity Nat'l Financial	Fidelity Nat'l Inf. Services	4.06	2
8	Hewlett-Packard	News Corporation	3.89	2
9	AutoZone	AutoNation, Inc.	3.8	3
10	United Tech. Corporation	PACCAR	3.74	2

CHAPTER 4

Fast Spatial Ancestry via Flexible Allele Frequency Surfaces

4.1 Abstract

Unique modeling and computational challenges arise in locating the geographic origin of people based on their genetic backgrounds. SNPs vary widely in informativeness, allele frequencies change nonlinearly with geography, and reliable localization requires evidence to be integrated across a multitude of SNPs. These problems become even more acute for people of mixed ancestry. It is hardly surprising that matching genetic models to computational constraints has limited the development of genographic estimation and projection. We attack these related problems by borrowing ideas from image processing and optimization theory. Our model pixelates the region of interest and operates SNP by SNP. We estimate allele frequencies across the landscape by maximizing a product of binomial likelihoods penalized by nearest neighbor interactions. Penalization smooths allele frequency estimates and promotes estimation at pixels with no data. Maximization is accomplished by an MM (minorize-maximize) algorithm. Once allele frequency surfaces are available, one can apply Bayes rule to compute the posterior probability that each pixel is the pixel of origin of a given person. Placement of admixed individuals on the landscape is more complicated and requires estimation of the fractional contribution of each pixel to a person's genome. This estimation problem also succumbs to a penalized MM algorithm. On the POPRES data, the current model gives better localization for both unmixed and admixed individuals than existing methods despite using just a small fraction of the available SNPs. Computing times are comparable to the best competing software.

4.2 Introduction

The pertinence of the first law of geography – "Everything is related to everything else, but near things are more related than distant things" [Tob70] – has long been obvious to population geneticists. For example, in the 1930's Fisher [Fis37, Fis00] and Kolmogorov [KPP37] derived and solved a partial differential equation describing the spatial spread of an advantageous allele. Subsequent generations of ecologists and evolutionary biologists [CTS00, SO78] have studied the correlations between geography and population structure from many different perspectives. During the last decade in particular, geneticists have discovered how to localize the origin of individuals, human and otherwise, based on their genetic backgrounds [LLN08, NJB08, WSC04, YNE12]. We will call such localization genographic projection.

In one of the more striking applications of principal component analysis (PCA), Novembre et al [NJB08] were able to match the first two principal components of the genotype matrix of the Population Reference Sample (POPRES) data set [NBK08] to the map of Europe. Genographic projection was accurate to within a few hundred kilometers. Though this level of resolution is impressive, it is natural to wonder if a model-based method of projection could perform better and whether inferences could be reliably made for admixed individuals. This prompted Yang et al [YNE12] to introduce spatial structure analysis (SPA). Not surprisingly, SPA produces more accurate genographic projections than PCA. In estimating allelic frequency surfaces for each surveyed SNP (single nucleotide polymorphism), SPA unfortunately depends on a simple linear parameterization of how allele frequencies vary with location. In practice, allele frequency surfaces can be bumpy without a dominant cline.

The current paper relaxes this restriction. Our software, SNPscape, adapts techniques from image reconstruction that encourage smooth but not strongly parameterized allele frequency surfaces [CS05, Lan90]. SNPscape is model based and fast. It can infer the geographic origin of Europeans in the POPRES data set to much less than 100 kilometers. Its impressive speed is achieved by focusing on the most informative markers, sometimes as few as 1% of all markers, and relying on new MM (minorization/maximization) algorithms for parameter estimation. In choosing ancestry informative markers, we replace the information

criterion of Rosenberg et al [RLW03] by a more sensible homogeneity likelihood ratio test that accommodates substantial differences in sample sizes.

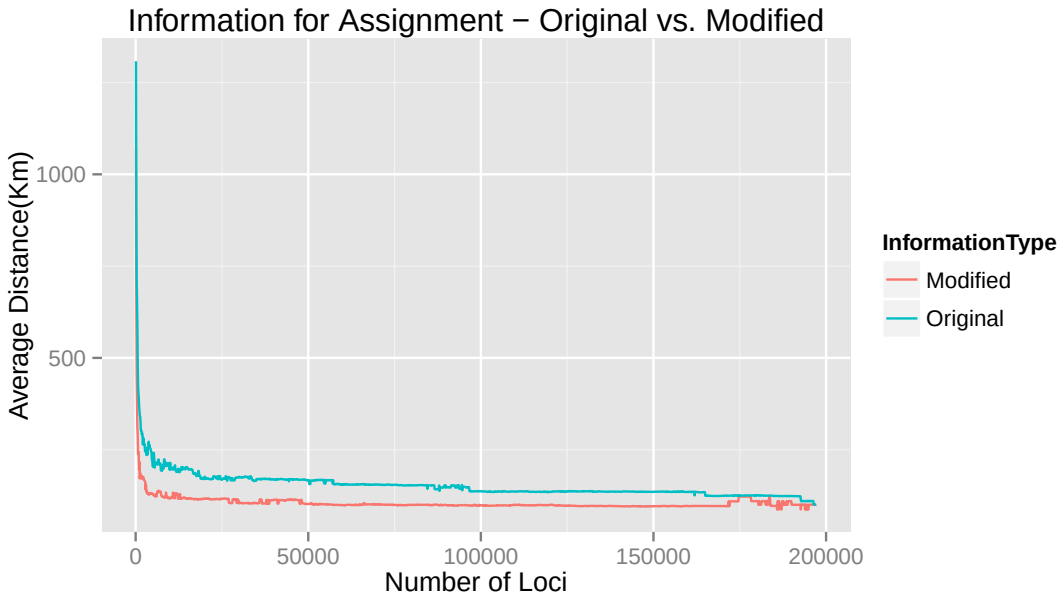


Figure 4.1: Average distance between the geographic origin of the POPRES individuals and their SNPscape estimated origins as a function of the number of SNPs employed. The figure reflects leave-one-out cross validation. The upper curve relies on the Rosenberg et al [RLW03] information criterion, and the lower curve relies on the LRT criterion.

4.3 Results

4.3.1 A Likelihood Ratio Criterion for SNP Selection

The utility of SNPs in identifying ancestral origins varies widely. The Rosenberg et al [RLW03] criterion for ranking SNPs makes the key assumption of equal sample sizes at the different sampling sites. In practice this assumption is usually violated. As an alternative, we turned to a likelihood ratio (LRT) statistic for testing homogeneity of allele frequencies across sites. The LRT statistic compares the best loglikelihood of the data under the null hypothesis of homogeneity to the best loglikelihood of the data under the alternative hy-

pothesis of complete heterogeneity. Figure 4.1 allows us to compare the value of the two different methods of ranking SNPs. The vertical axis of the figure represents the average distance under cross validation between the true location of the POPRES individuals and their estimated locations under SNPscape. The horizontal axis represents the number of SNPs employed, with SNPs taken in their order of informativeness. Although both curves document the value of ancestry informative SNPs in geographical projection, it is obvious that the likelihood ratio criterion performs better than the information criterion.

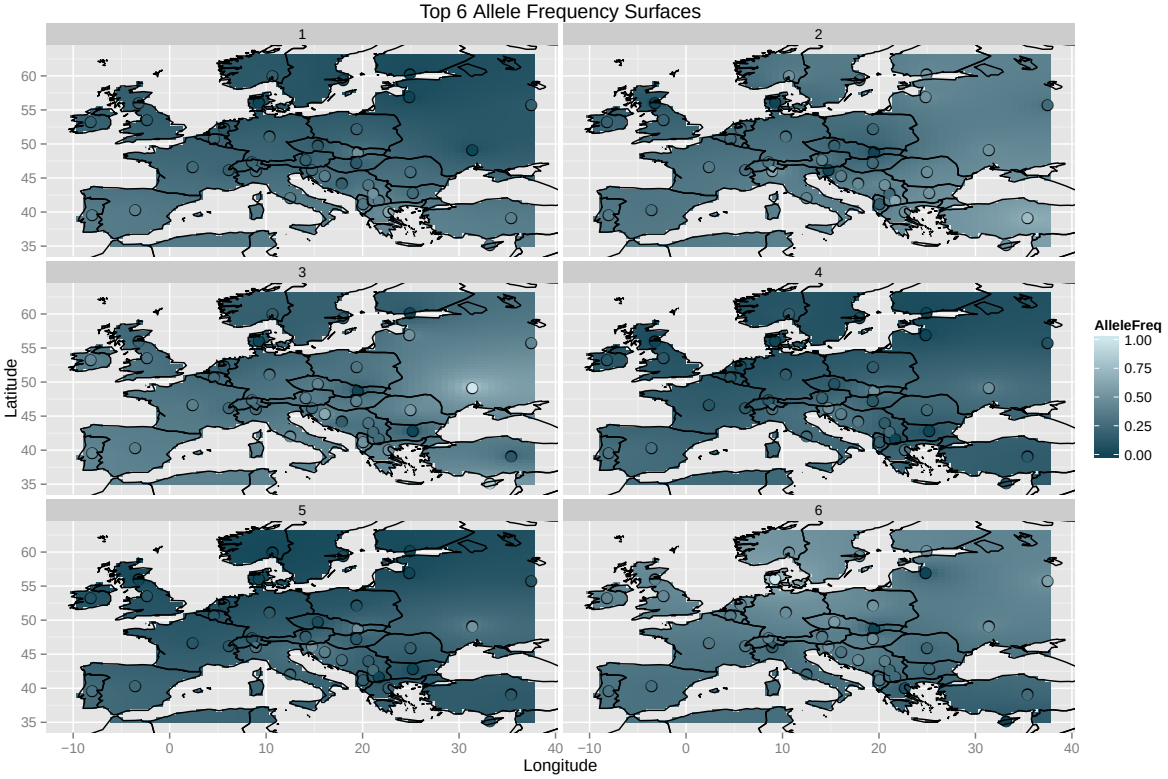


Figure 4.2: Allele frequency surfaces generated by SNPscape with tuning parameter $\rho = 0.1$ for the six most informative SNPs. These surfaces are overlaid with filled in circles conveying the MLE estimates for each sampled site. For the sake of comparison, the same SNP surfaces are depicted in Figure 4.3 for SPA.

4.3.2 Allele Frequency Surfaces

Accurate allele frequency surfaces are the primary reason for SNPscape's superior performance. SNPscape surfaces are more adaptable and less rigidly parameterized. Figure 4.2 depicts the estimated allele frequency surfaces of the six most informative SNPs of the POPRES data. The figure also plots the maximum likelihood estimates for each sampled site as a filled in circle at the appropriate location. For comparison, Figure 4.3 depicts the surfaces for the same SNPs generated by SPA. The figures demonstrate that SNPscape surfaces match the sampled circles better than the SPA surfaces. SPA appears to be too heavily influenced by outlier sites and less adaptable overall.

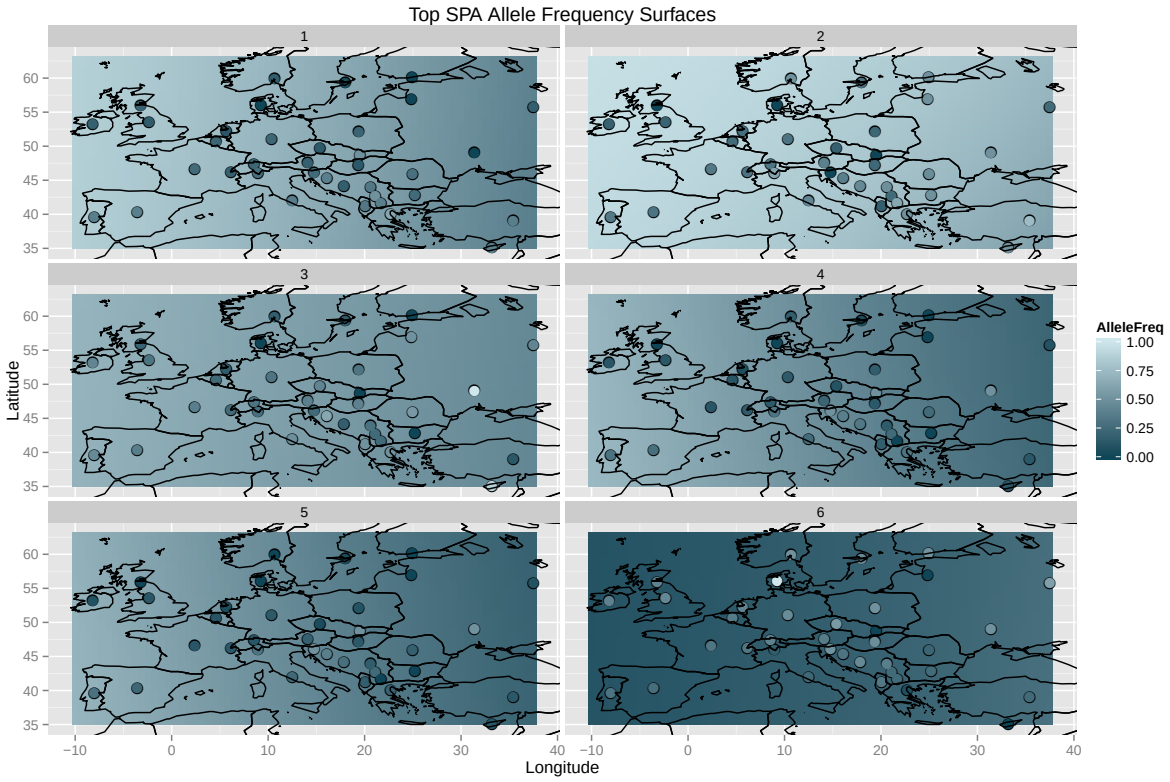


Figure 4.3: Allele frequency surfaces generated by SPA for the six most informative SNPs. These surfaces are overlaid with filled in circles conveying the MLE estimates for each sampled site. For the sake of comparison, the same SNP surfaces are depicted in Figure 4.2 for SNPscape.

4.3.3 Ancestral Origin Inference

Genographic projection is the main application of SNPscape. To showcase SNPscape's accuracy, we computed average localization error by leave-one-out cross validation. Figure 4.4 displays the results for SNPscape versus SPA. The lower curve for SPA emphasizes the benefits of exploiting LRT ordered SNPs. Examination of the figure shows that SNPscape using 1% of the SNPs achieves better accuracy than SPA using all of the SNPs. Using 5% of the SNPs, SNPscape is nearly perfect at the pixel level in its localizations. The same point can be made by comparing SNPscape's results to the results in Table 4.1 of the SPA paper [YNE12]. Given the nature of the table, a fair comparison requires using SNPscape to estimate the optimal ancestral origin of each person and then assigning the person to the closest sampling site as measured by geodesic distance. Overall, SNPscape was more than twice as accurate as PCA and SPA based on just 1% of the data. With 5% of the data, SNPscape maps individuals to sampled sites with 99% accuracy.

We also performed some small-scale comparisons to SCAT [WSC04]. SCAT is so slow that more ambitious comparisons are impossible. Table 4.2 records the average distance to the true origin and the run times for the 100 most informative SNPs. SNPscape and SCAT place individuals almost 2000Km closer to their true origin than SPA. SPA is the fastest (1 minute) of the three programs, followed closely by SNPscape (2 minutes) and distantly by SCAT (362 minutes). Thus, the current version of SNPscape delivers good placement with competitive execution times. SCAT is ill equipped to handle large numbers of SNPs.

4.3.4 Estimating Proportions of Admixed Origins

Many individuals have mixed ancestry. PCA tends to localize individuals with parents of different ethnicities in between their parents' regions of origin. SPA has the capacity to localize each parent separately, but the user must inform the program beforehand how many different ancestries contribute to a given individual. Because this information is often unavailable, it would be preferable for admixture detection and origin selection to be more agnostic. SNPscape can estimate admixture fractions on a pixel by pixel basis. For example,

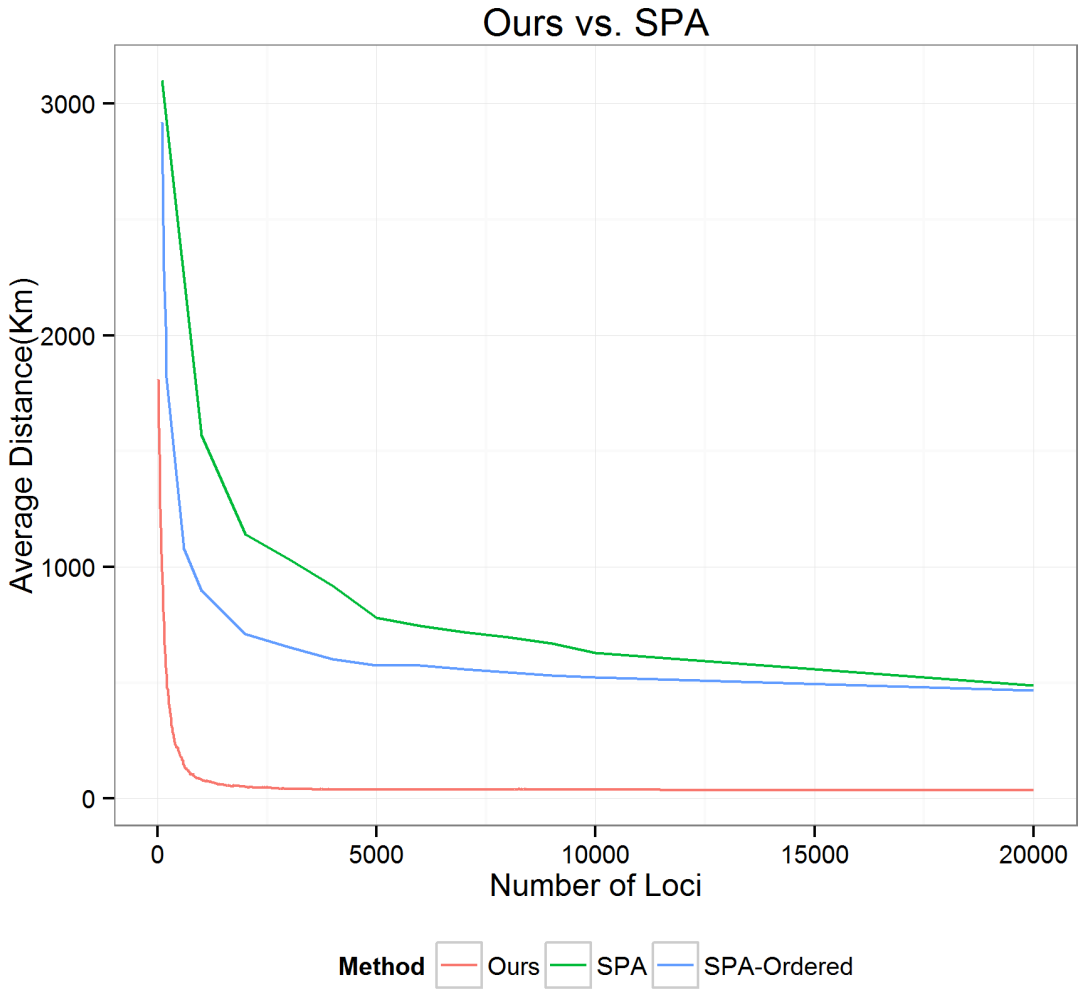


Figure 4.4: Average localization error for individuals based on leave-one-out cross validation using SNPscape ($\rho = 0.1$), SPA without SNP selection, and SPA with SNP selection.

when applied to a person with a German parent and an Italian parent, ideally SNPscape should deliver 50% German ancestry, 50% Italian ancestry, and 0% Swedish, Russian, and other ancestry. This would make SNPscape comparable to the program ADMIXTURE [ANL09], with the benefit of using more accurate allele frequencies and encompassing small countries with no sampled people at all.

The methods section describes our model for admixture estimation with SNPscape. Table 4.3 presents results for simulated individuals whose parents originate at different sites.

Table 4.1: Comparison of localization by population. Population of origin was predicted for each individual using leave-one-out cross validation. Accuracy ± s.d. is the proportion of individuals from each population correctly assigned to their true population. The values listed for SNPscape represent either 1% of the data (2K SNPs) or 5% of the data (10K SNPs). To make the values from SNPscape comparable to PCA and SPA, the most likely location of each individual was estimated, and the population closest in distance to that point was chosen as the population of origin. The results for PCA and SPA are taken from Table 1 of the paper [YNE12].

		Accuracy			
				SNPscape	
Geographic Origin	Num. Individuals	PCA	SPA	1% of Data	5% of Data
Italy	219	0.70 ± 0.03	0.74 ± 0.03	0.99 ± 0.01	0.99 ± 0.01
UK	200	0.44 ± 0.04	0.53 ± 0.04	1.00 ± 0.00	1.00 ± 0.00
Spain	136	0.71 ± 0.04	0.69 ± 0.04	0.98 ± 0.01	0.99 ± 0.01
Portugal	128	0.20 ± 0.04	0.38 ± 0.04	0.98 ± 0.01	1.00 ± 0.00
Switzerland-French	125	0.26 ± 0.04	0.33 ± 0.04	0.95 ± 0.02	0.97 ± 0.02
France	89	0.70 ± 0.05	0.66 ± 0.05	0.95 ± 0.02	0.97 ± 0.02
Switzerland-German	84	0.23 ± 0.05	0.27 ± 0.05	1.00 ± 0.00	0.99 ± 0.01
Germany	71	0.25 ±0.05	0.28 ± 0.05	1.00 ± 0.00	1.00 ± 0.00
Ireland	61	0.28 ± 0.06	0.28 ± 0.06	0.92 ± 0.03	1.00 ± 0.00
Yugoslavia	44	0.25 ± 0.07	0.30 ± 0.07	1.00 ± 0.00	1.00 ± 0.00
Mean	115.7	0.40 ± 0.05	0.45 ± 0.05	0.98 ± 0.01	0.99 ± 0.01

In admixture mode SNPscape exploits the same allele frequency surfaces that it does in normal mode. However, instead of applying Bayes' rule to find the posterior probability of origin of each pixel, it estimates an admixture fraction for each pixel by penalized maximum likelihood estimation. Surprisingly, SNPscape is not only able to select the two contributing populations, it is also able to estimate their proportions well. Overall, it attributes an average of 0.43 ± 0.07 to the first country of origin and 0.39 ± 0.08 to the second country origin, leaving an average of 18% misattribution. For reference, we have also included the nearly identical results from ADMIXTURE. In Figure 4.5 and 4.6, we took the admixture estimation a step further by estimating the fractions at each pixel instead of each population. This

Table 4.2: Accuracy of origin localization and run times for SNPscape, SCAT, and SPA for 100 SNPs. SPA's localizations can fall outside the mapped region. The localization of SNPscape and SCAT must fall within the mapped region.

Method	Placement(Km)	Time(m)
SNPscape	981	2
SCAT	1074	362
SPA	2920	1

analysis allows us to place individuals at locations with no sampled data. In the figure the true locations of the individual's parents are highlighted in white, while SNPscape's results are written as text at their respective locations. All values below 1% are omitted.

Table 4.3: Comparison of SNPscape restricted to population pixels and Admixture. Based on simulated genotypes of admixed individuals, the table displays true populations along with the average estimated proportion of the genome coming from each of the true populations. The results rely on 20K SNPs and 5 individuals with parents from each pair of populations.

		SNPscape		Admixture	
Origin 1	Origin 2	Fraction 1	Fraction 2	Fraction 1	Fraction 2
Italy	UK	0.45 ± 0.11	0.33 ± 0.08	0.45 ± 0.11	0.33 ± 0.08
Italy	Portugal	0.38 ± 0.10	0.44 ± 0.08	0.38 ± 0.10	0.44 ± 0.08
Italy	Spain	0.45 ± 0.05	0.39 ± 0.08	0.45 ± 0.05	0.39 ± 0.08
Switzerland-French	UK	0.42 ± 0.04	0.40 ± 0.08	0.42 ± 0.04	0.40 ± 0.08
Portugal	UK	0.40 ± 0.03	0.41 ± 0.05	0.40 ± 0.03	0.41 ± 0.05
Spain	UK	0.41 ± 0.09	0.33 ± 0.04	0.41 ± 0.09	0.34 ± 0.04
Portugal	Spain	0.45 ± 0.07	0.44 ± 0.09	0.45 ± 0.07	0.44 ± 0.09
France	Italy	0.43 ± 0.08	0.38 ± 0.11	0.43 ± 0.08	0.38 ± 0.11
Germany	Italy	0.38 ± 0.11	0.42 ± 0.10	0.38 ± 0.11	0.41 ± 0.10
Germany	Portugal	0.48 ± 0.05	0.38 ± 0.07	0.48 ± 0.05	0.38 ± 0.07
	Mean	0.43 ± 0.07	0.39 ± 0.08	0.43 ± 0.07	0.39 ± 0.08

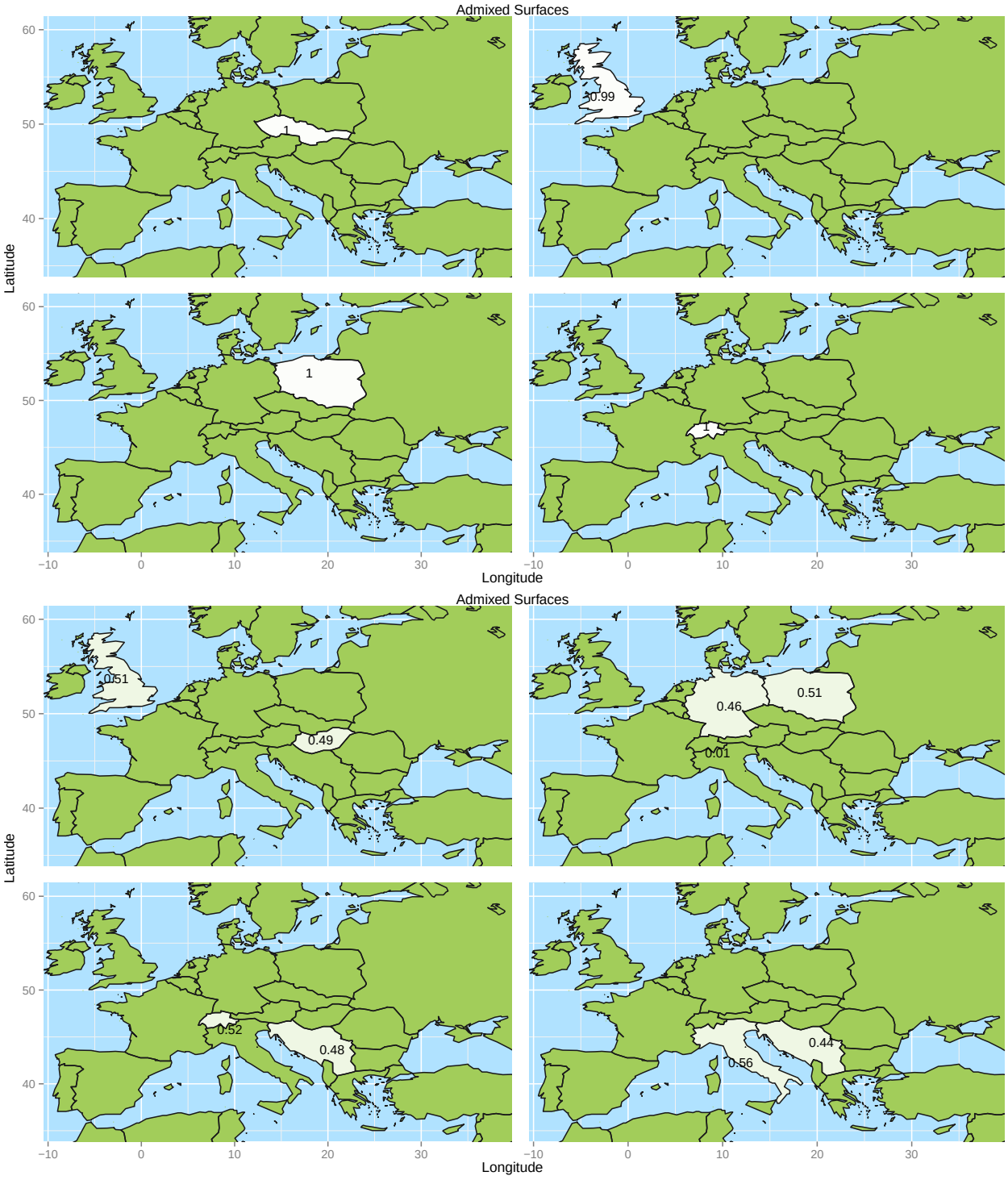


Figure 4.5: Admixture coefficients for four simulated Europeans. The true locations of an individual's grandparents are highlighted in white and may overlap. SNPscape's admixture coefficients are printed as fractions at their respective locations. Values below 1% are omitted.

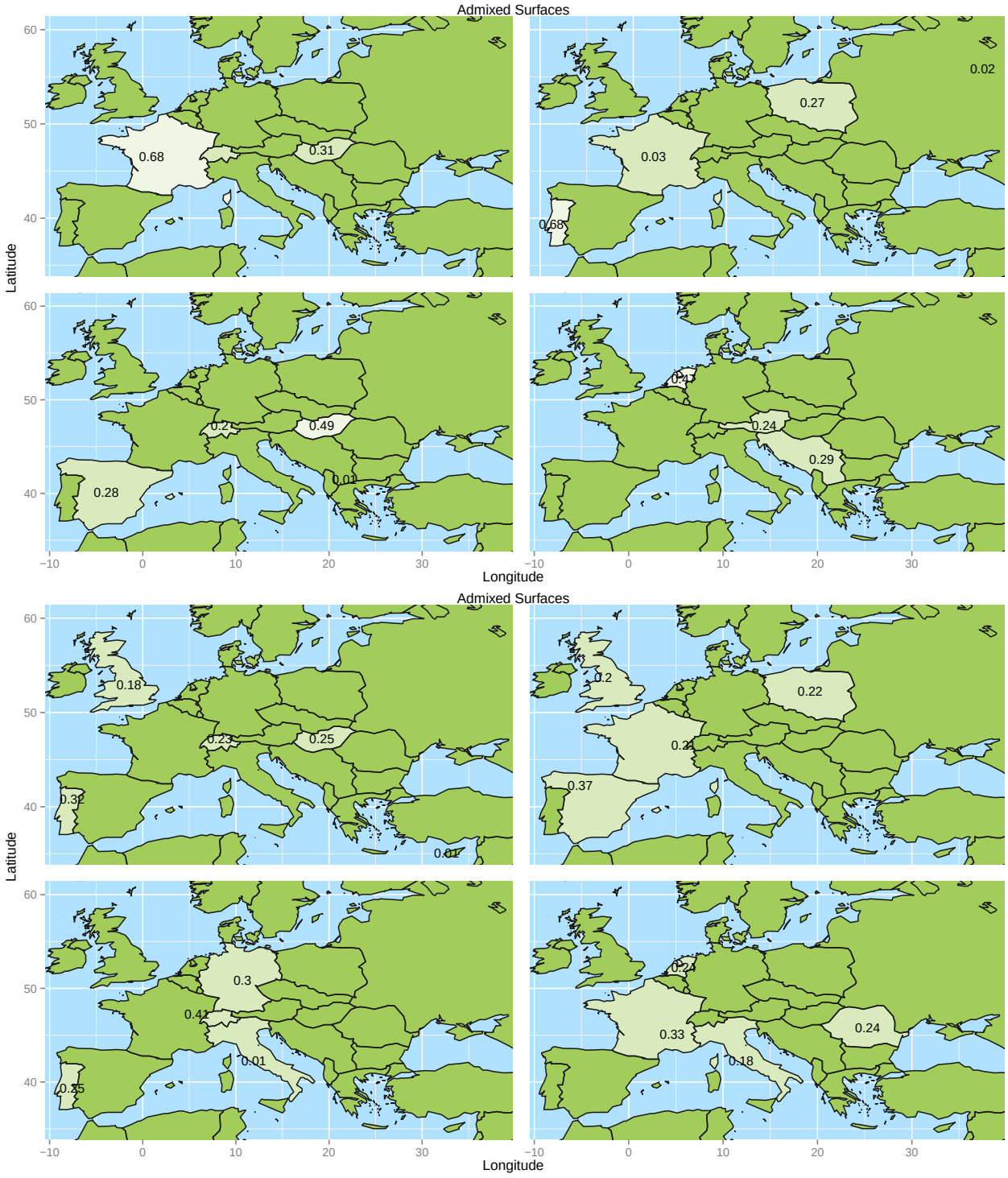


Figure 4.6: Admixture coefficients for four simulated Europeans. The true locations of an individual's grandparents are highlighted in white and may overlap. SNPscape's admixture coefficients are printed as fractions at their respective locations. Values below 1% are omitted.

4.4 Methods

4.4.1 A Likelihood Ratio Criterion for SNP Selection

The majority of SNPs are uninformative for ancestry and geographic localization. This fact and considerations of computational speed suggest choosing the most informative SNPs and ignoring the rest. The standard ancestry informativeness criterion of Rosenberg et al [RLW03] makes the implicit assumption of equal sample sizes. The failure of this assumption in the POPRES data prompted us to turn to a homogeneity likelihood ratio test. The null model of the test for a given SNP postulates that all individuals come from a single population with a unique allele frequency for the reference allele; the alternative model postulates different reference allele frequencies at the different sampling sites. Binomial sampling is in force. Suppose there are s sites with k_i sampled reference alleles and n_i sampled genes (reference alleles plus alternative alleles) at site i. If $k = \sum_{i=1}^{s} k_i$, and $n = \sum_{i}^{s} n_i$, then the likelihood ratio statistic reduces to

$$LRT \;=\; 2\ln \frac{\max_{\boldsymbol{p}} \prod_{i=1}^{s} \binom{n_i}{k_i} p_i^{k_i}(1-p_i)^{n_i-k_i}}{\prod_{i=1}^{s} \binom{n_i}{k_i} \max_q q^k(1-q)^{n-k}} \;=\; 2\ln \frac{\prod_{i=1}^{s} \hat{p}_i^{k_i}(1-\hat{p}_i)^{n_i-k_i}}{\hat{q}^k(1-\hat{q})^{n-k}},$$

where $\hat{q} = k/n$ and $\hat{p}_i = k_i/n_i$ are the maximum likelihood estimates of the reference allele frequencies under the null and alternative models, respectively. Although small sample sizes at many sites invalidate the chi-square distribution of the likelihood ratio statistic, nothing prevents the statistic from being used as an index to rank the various SNPs. In our experience, the highest ranking SNPs are indeed the most informative.

4.4.2 Allele Frequency Surface Estimation

To estimate the allele frequency surface for a given SNP, we pixelate the region of interest, say Europe, and assign a reference allele frequency p_i to each pixel i. Extending our previous notation, k_i now represents the number of sampled reference alleles and n_i sampled genes from pixel i. For most pixels, $k_i = n_i = 0$. Maximizing the binomial loglikelihood

$$L(\mathbf{p}) \;=\; \sum_i \left[\ln \binom{n_i}{k_i} + k_i \ln p_i + (n_i - k_i)\ln(1-p_i)\right]$$

would allow estimation of the reference allele frequencies if all pixels actually contained sampled people. Because this is not the case and because we desire smooth estimates across the landscape, we subtract squared difference penalties from the loglikelihood and maximize the penalized loglikelihood

$$\begin{aligned} f(\mathbf{p}) &= L(\mathbf{p}) - \rho \sum_{\{i,j\}} w_{ij}(p_i - p_j)^2 \\ &= \sum_i \left[\ln \binom{n_i}{k_i} + k_i \ln p_i + (n_i - k_i) \ln(1 - p_i) \right] - \rho \sum_{\{i,j\}} w_{ij}(p_i^2 - 2p_i p_j + p_j^2) \end{aligned} \tag{4.1}$$

Here the tuning constant ρ determines the extent of smoothing. The nonnegative weights w_{ij} incorporate nearest neighbor interactions and scale the distance between pixel centers. For square pixels we accordingly set $w_{ij} = 1$ for pixels sharing a side, $w_{ij} = 1/\sqrt{2}$ for pixels sharing a corner, and $w_{ij} = 0$ for all other pixel pairs. Limiting interactions to the eight pixels surrounding a pixel obviously reduces computational complexity.

Maximizing the criterion (4.1) is a formidable optimization problem because the penalty terms couple the parameters in an awkward fashion and make it impossible to find an exact solution. However, we can invoke the MM (minorization-maximization) principle [HL04, Lan12, LHY00] and construct a surrogate function that separates the parameters. A surrogate function $g(\boldsymbol{p} \mid \boldsymbol{p}_n)$ for the objective function $f(\boldsymbol{p})$ must be tangent to $f(\boldsymbol{p})$ at the current iterate $\boldsymbol{p}_n$ and dominated by it throughout the common domain of both functions. Formally, these conditions amount to the equality $g(\boldsymbol{p}_n \mid \boldsymbol{p}_n) = f(\boldsymbol{p}_n)$ and the inequality $g(\boldsymbol{p} \mid \boldsymbol{p}_n) \leq f(\boldsymbol{p})$ for all feasible $\boldsymbol{p}$. In the maximization step of the MM algorithm, the next iterate $\boldsymbol{p}_{n+1}$ is chosen to maximize $\boldsymbol{p} \mapsto g(\boldsymbol{p} \mid \boldsymbol{p}_n)$. These definitions imply the ascent condition

$$f(\boldsymbol{p}_{n+1}) \leq g(\boldsymbol{p}_{n+1} \mid \boldsymbol{p}_n) \leq g(\boldsymbol{p}_n \mid \boldsymbol{p}_n) = f(\boldsymbol{p}_n),$$

which is the secret of the MM principle's success.

The derivation of our surrogate function depends on the minorization

$$xy \geq x_m y_m \left[1 + \ln\left(\frac{x}{x_m}\right) + \ln\left(\frac{y}{y_m}\right) \right]$$

for positive numbers. This minorization reduces to the supporting line inequality $-\ln z \geq 1 - z$ if we substitute $z = xy/(x_n y_n)$. Application of this majorization leads to the minorization

$$g(\boldsymbol{p} \mid \boldsymbol{p}_n) = L(\boldsymbol{p}) - \rho \sum_{\{i,j\}} w_{ij} \left[p_i^2 + p_j^2 - 2p_{ni}p_{nj}(\ln p_i + \ln p_j) \right]$$

of $f(\boldsymbol{p})$ up to an irrelevant constant. With parameters separated, we can now solve the stationarity equation

$$0 = \frac{k_i}{p_i} - \frac{n_i - k_i}{1 - p_i} - 2\rho \sum_{j \neq i} w_{ij} \left[p_i - p_{ni}p_{nj}\frac{1}{p_i} \right]$$

for the MM update $p_{n+1,i}$ of p_i. Multiplying this equation by $p_i(1 - p_i)$ yields an equivalent cubic polynomial equation

$$0 = a_i p_i^3 - a_i p_i^2 - (n_i + b_{ni})p_i + k_i + b_{ni},$$

where $a_i = 2\rho \sum_{j \neq i} w_{ij}$ and $b_{ni} = 2\rho p_{ni} \sum_{j \neq i} w_{ij} p_{nj}$. This cubic is positive when $p_i = 0$ and nonpositive when $p_i = 1$. The cubic also tends to $\pm\infty$ when p_i tends to $\pm\infty$. Hence, there exists a single root on the interval $(0, 1]$. One can extract this root by one of the standard formulas for solving a cubic equation.

4.4.3 Localization of Unknowns

Once the allele frequency surfaces for the informative SNPs are estimated by the MM algorithm, one can localize individuals of unknown origin. For person j with genotype vector $\boldsymbol{x}_j$, Bayes' rule gives the posterior probability

$$\Pr(j \text{ from pixel } i \mid \boldsymbol{x}_j) = \frac{\Pr(\boldsymbol{x}_j \mid j \text{ from pixel } i) \Pr(j \text{ from pixel } i)}{\Pr(\boldsymbol{x}_j)}$$

that j originates from pixel i. Application of this rule depends on fixing a prior. Two possibilities are convenient. The simpler one is the uniform prior. A more accurate but less convenient choice is to scale the prior of a pixel by its population size. For sufficiently informative genetic data, the evidence dominates the prior, and the uniform prior is probably adequate. The likelihood term

$$\Pr(\boldsymbol{x}_j \mid j \text{ from pixel } i) = \prod_k \Pr(x_{jk} \mid j \text{ from pixel } i)$$

factors into a product of likelihoods at the canvassed SNPs under the assumption of linkage and Hardy-Weinberg equilibrium. The likelihood $\Pr(x_{jk} \mid j \text{ from pixel } i)$ at SNP k equals one of the three genotype probabilities p_{ik}^2, $2p_{ik}(1-p_{ik})$, or $(1-p_{ik})^2$ depending on j's genotype at SNP k. In practice, it is advisable to work with the logarithms of these quantities to avoid computer underflows. Although the pixel with the highest posterior probability provides the most likely localization, it is a good idea in practice to assign an average latitude and longitude and highlight the set of pixels that contribute substantially to the posterior distribution.

4.4.4 Admixed Individuals

For an ethnically admixed individual, we suggest estimating the fractional contribution f_i of each pixel i to his/her genome. If we let x_k denoted the observed number of reference alleles at SNP k, then the loglikelihood of the person's observed genotypes amounts to

$$L(\mathbf{f}) \;=\; \sum_k \Big\{ x_k \ln\Big(\sum_i f_i p_{ik}\Big) + (2-x_k)\ln\Big[\sum_i f_i(1-p_{ik})\Big]\Big\}$$

subject to the constraints $\sum_i f_i = 1$ and $f_i \geq 0$ for all i. This formulation of the problem is reminiscent of the ethnic admixture problem if we identify pixels with ethnic groups and fix allele frequencies rather than estimate them [AL11, ANL09]. Maximization of $L(\mathbf{f})$ is a typical MM exercise. The key step is to separate parameters via the minorizations

$$\begin{aligned}
\ln\Big(\sum_i f_i p_{ik}\Big) &\geq \sum_i \frac{f_{ni}p_{ik}}{\sum_j f_{nj}p_{jk}} \ln\Big(\frac{\sum_j f_{nj}p_{jk}}{f_{ni}p_{ik}} f_i p_{ik}\Big) \\
\ln\Big[\sum_i f_i(1-p_{ik})\Big] &\geq \sum_i \frac{f_{ni}(1-p_{ik})}{\sum_j f_{nj}(1-p_{jk})} \ln\Big[\frac{\sum_j f_{nj}(1-p_{jk})}{f_{ni}(1-p_{ik})} f_i(1-p_{ik})\Big]
\end{aligned}$$

based on Jensen's inequality applied to the concave function $\ln x$. Equality holds when $\boldsymbol{f} = \boldsymbol{f}_n$. If we define the constants

$$c_{nik} \;=\; x_k \frac{f_{ni}p_{ik}}{\sum_j f_{nj}p_{jk}} + (2-x_k)\frac{f_{ni}(1-p_{ik})}{\sum_j f_{nj}(1-p_{jk})},$$

then standard arguments invoked in maximizing a multinomial likelihood yield the updates

$$f_{n+1,i} \;=\; \frac{\sum_k c_{nik}}{\sum_k \sum_j c_{njk}}.$$

One can accelerate convergence in estimating $\boldsymbol{f}$ and improve inference by imposing a penalty that drives to 0 those components f_i with low explanatory power. Since a lasso penalty $\lambda \sum_i f_i = \lambda$ is effectively constant, we suggest the alternative penalty $\lambda \sum_i q(f_i)$, where the penalty function

$$q(f) = \begin{cases} f & f < \delta \\ \delta & f \geq \delta \end{cases}$$

relies on a positive threshold δ beyond which no further penalty is imposed. Because $q(f)$ is nondifferentiable, we minorize it by the linear function f on the domain $[0, \delta)$ and by the constant δ on the domain $[\delta, \infty)$. We previously employed a variant of the current admixture model and penalty to estimate haplotype frequencies. Rather than repeat the mathematical derivation of the same penalized MM algorithm here, we refer the reader to the reference [AL08] for details. Suffice it to say that with parameters separated, the MM updates require solving a simple quadratic equation for each component f_i. Generic extrapolation techniques for MM and similar algorithms permit convergence acceleration beyond that afforded by our specific penalization [ZAL11].

4.5 Discussion

Motivated by advances in image reconstruction, we present a probability model for the estimation of complex allele frequency surfaces. Our model can capture not only linear clines, but also multiple local peaks on a pixelated landscape. Allele frequency estimates represent a compromise between locally sampled genotypes and smoothness. The degree of smoothness is determined empirically by cross validation. Genographic projection exploits the allele frequency surfaces of the most informative SNPs. To no one's surprise, the ancestry informative SNPs drive projection. In ranking SNPs our homogeneity likelihood ratio statistic outperforms the information criterion of Rosenberg et al [RLW03], which assumes equal sample sizes at the sampled sites. In practice, the combination of a good model with just 1% of the available SNPs give better geographic localization on the POPRES data than competing models (PCA, SCAT, and SPA) with all of the SNPs. Our computing times are

vastly superior to SCAT and competitive with PCA and SPA.

We have also proposed a model for genographic projection of admixed individuals. Our model assigns an admixture coefficient to each pixel. To avoid over-parameterization, we impose a penalty that enforces parsimony and focuses attention on those pixels with the greatest explanatory power. Estimation of both allele frequency surfaces and admixture coefficients benefits from the MM principle. The MM algorithms generated are simple to code and automatically enjoy the ascent property. Convergence can be slow, but standard extrapolation techniques accelerate convergence dramatically. On the negative side of the balance sheet, our software SNPscape requires more storage per SNP than SPA, which characterizes an allele frequency surface by just three parameters. The accuracy of SNPscape is also limited by pixelation. For example, in our analysis of the POPRES dataset, we chose a 70×70 grid of pixels. This gives a center to center distance of 61 Km for diagonally adjacent pixels and a resolution of 30.5 Km at best. A finer pixelation would improve matters at the expense of more computation. The heat maps of posterior probabilities and admixture coefficients afforded by pixelation are a decided plus. The ability to exclude infeasible pixels over oceans is another advantage.

Modeling is an art. The best models combine realism with computational efficiency. The injection of ideas and techniques from image reconstruction is a major contribution of SNPscape. Pixelation and nearest neighbor interactions offer a logical framework for estimation. MM algorithms are also ubiquitous in imaging. Our admixture model is more directly motivated by genetic considerations. It cleanly circumvents the need for specifying which ancestors of an admixed person should be taken as geographically localized. Finally, our SNP selection criterion is probably better suited to identifying ancestry informative SNPs than abstract information criterion. Readers will doubtless think of many other ways of improving the current model. Science, like product design, is usually an iterative process of successive refinement.

4.6 Supplementary Results

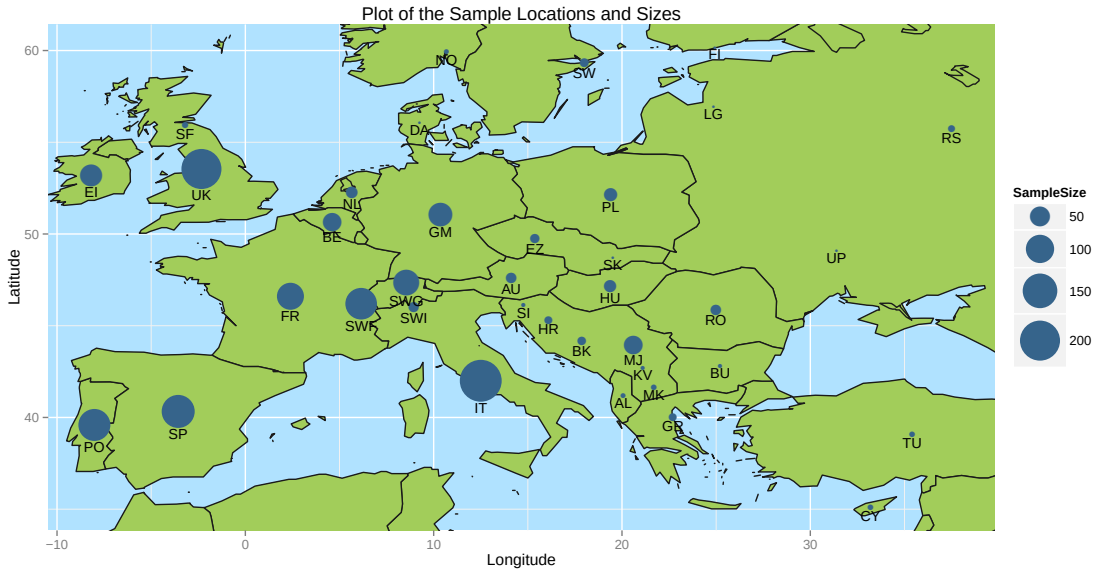


Figure 4.7: A plot of the locations and sample sizes of the POPRES dataset.

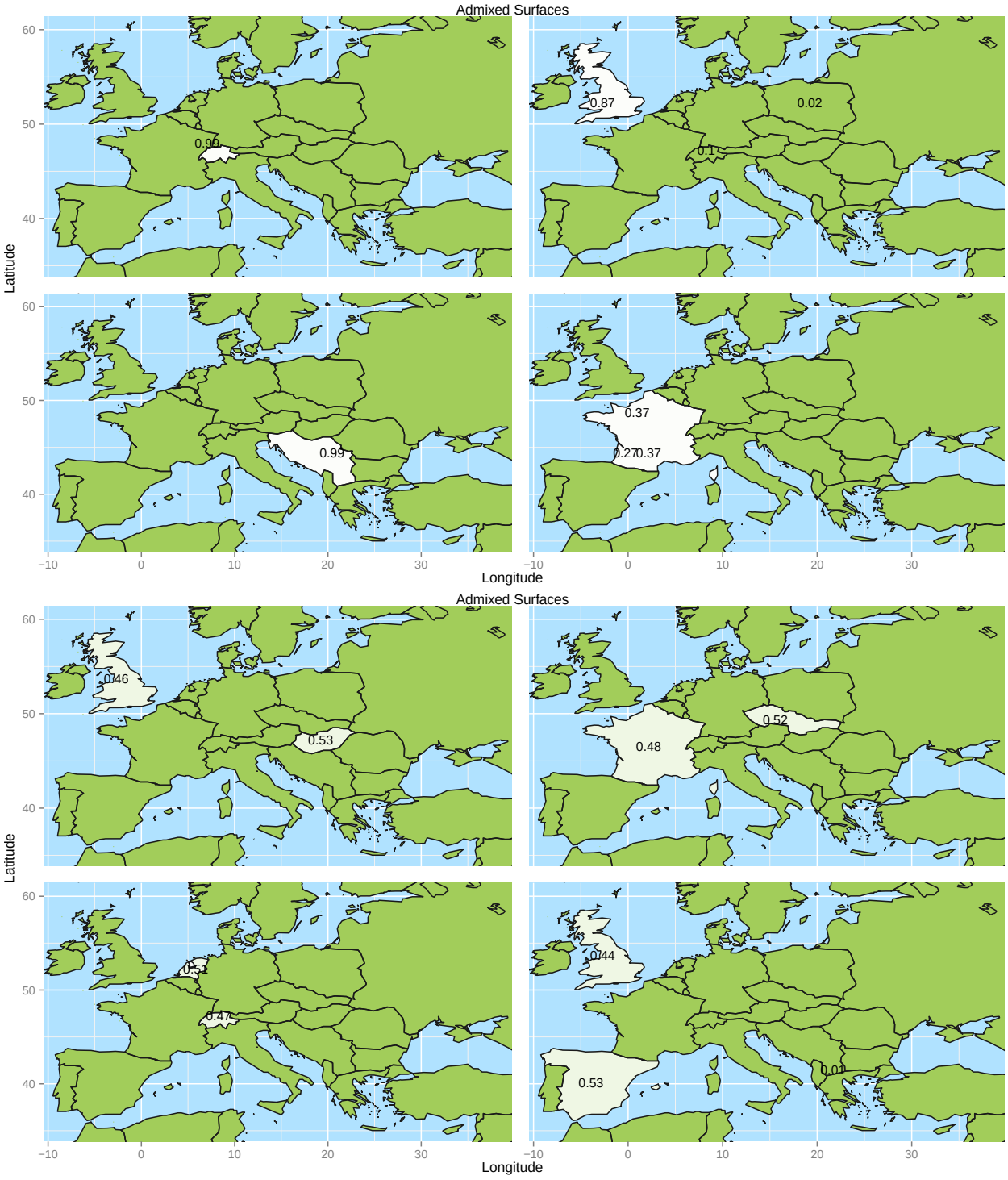


Figure 4.8: Additional admixture coefficients for four simulated Europeans. The true locations of an individual's grandparents are highlighted in white and may overlap. SNPscape's admixture coefficients are printed as fractions at their respective locations. Values below 1% are omitted.

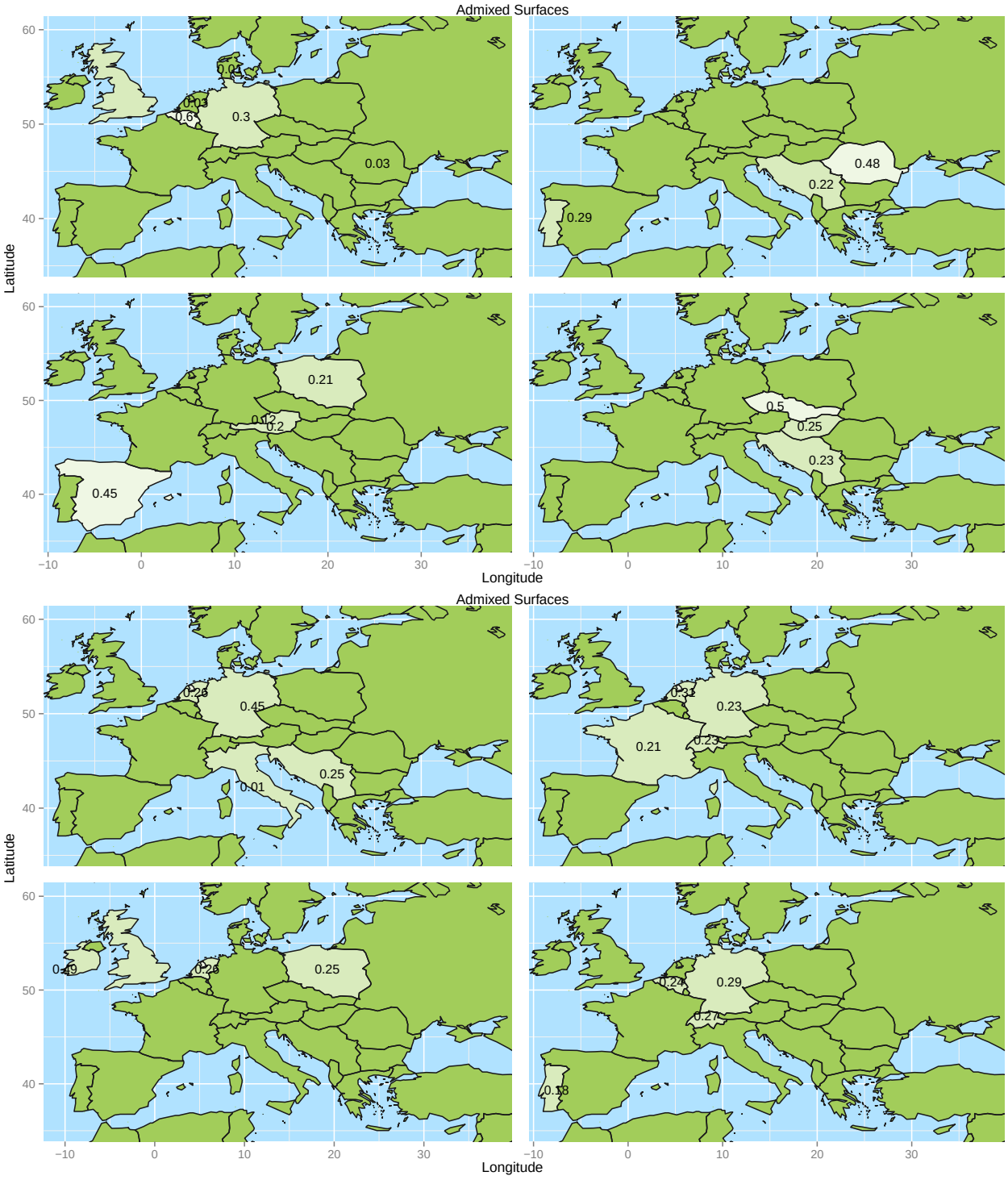


Figure 4.9: Additional admixture coefficients for four simulated Europeans. The true locations of an individual's grandparents are highlighted in white and may overlap. SNPscape's admixture coefficients are printed as fractions at their respective locations. Values below 1% are omitted.

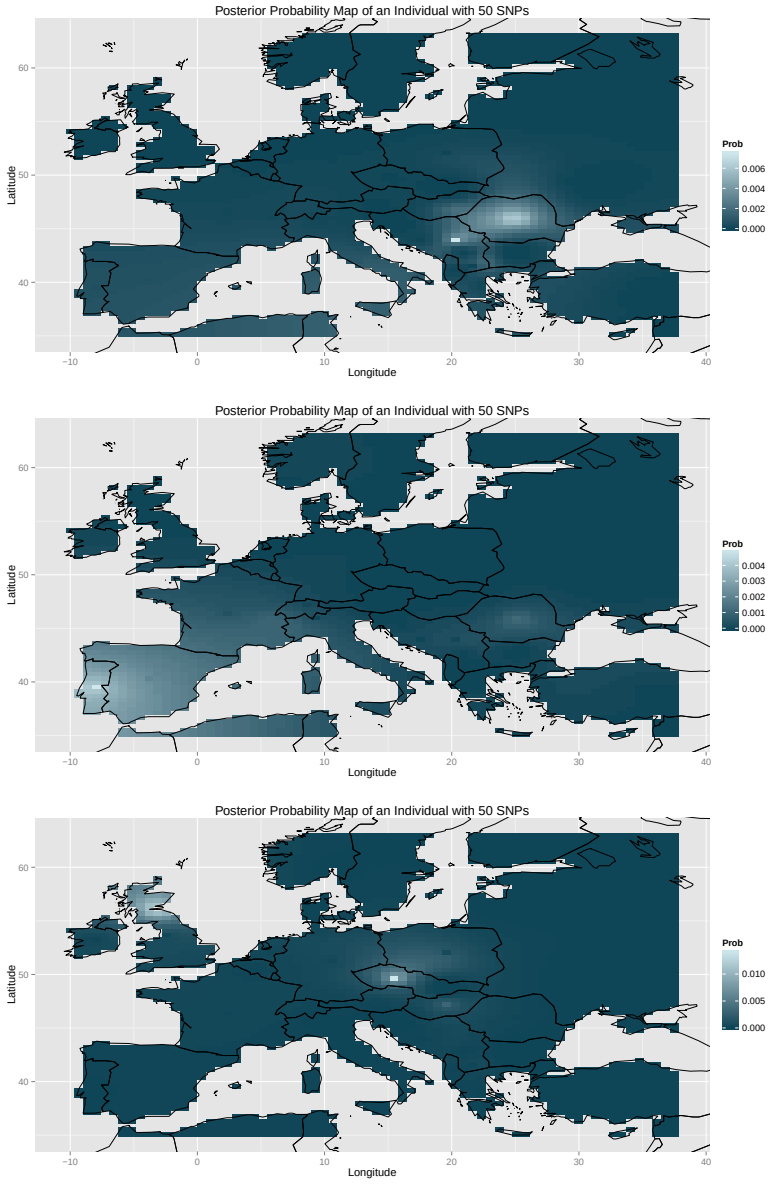


Figure 4.10: Plot of the posterior probability of a three different individuals coming from each pixel using 50 SNPs with $\rho = 0.1$.

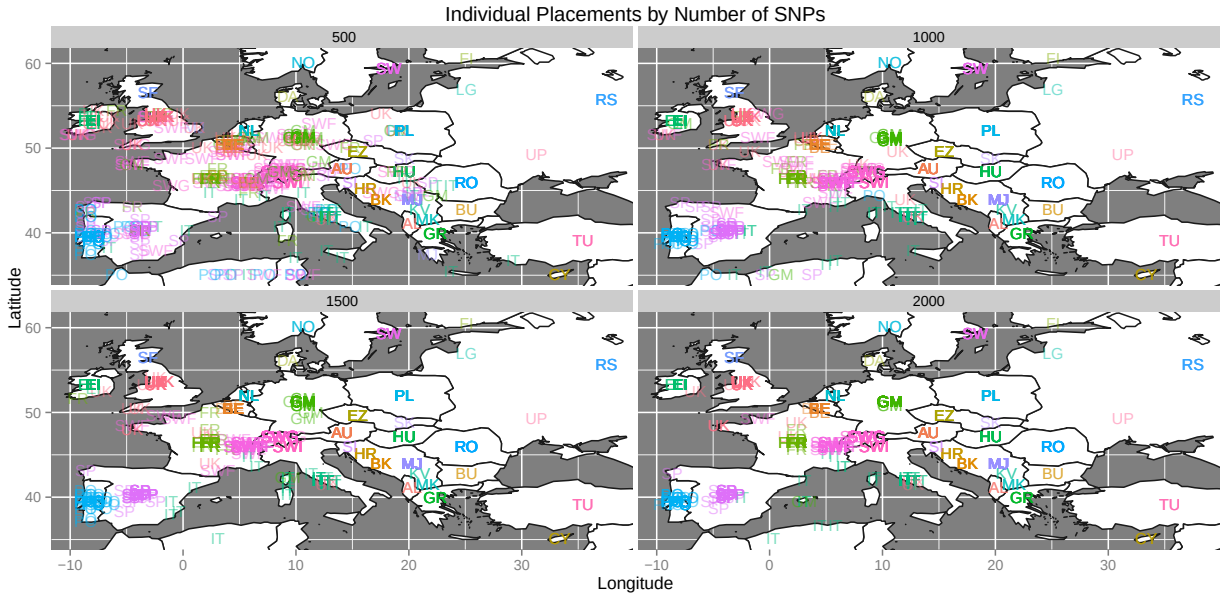


Figure 4.11: This figure shows the placement of all individuals back onto the map after using their data to generate allele frequency surfaces using various numbers of SNPs with $\rho = 0.1$. Note that the text is somewhat transparent to allow visualization of individuals directly on top of each other. For example there are many individuals placed in Russia(RS) at the same spot.

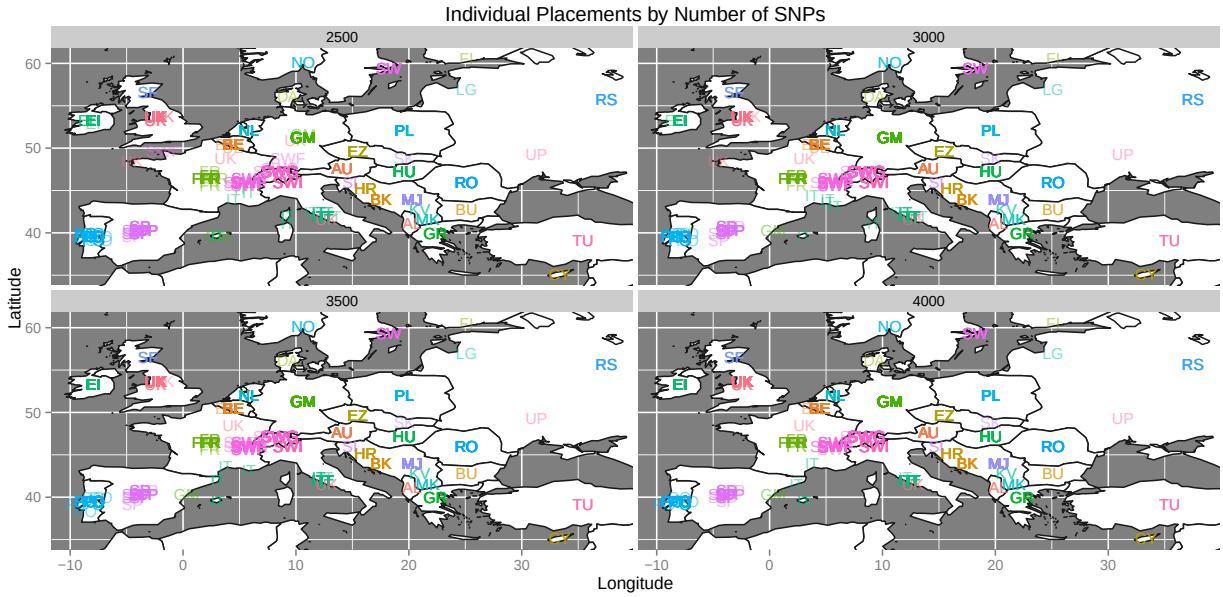


Figure 4.12: This figure shows the placement of all individuals back onto the map after using their data to generate allele frequency surfaces using various numbers of SNPs with $\rho = 0.1$. Note that the text is somewhat transparent to allow visualization of individuals directly on top of each other. For example there are many individuals placed in Russia(RS) at the same spot.

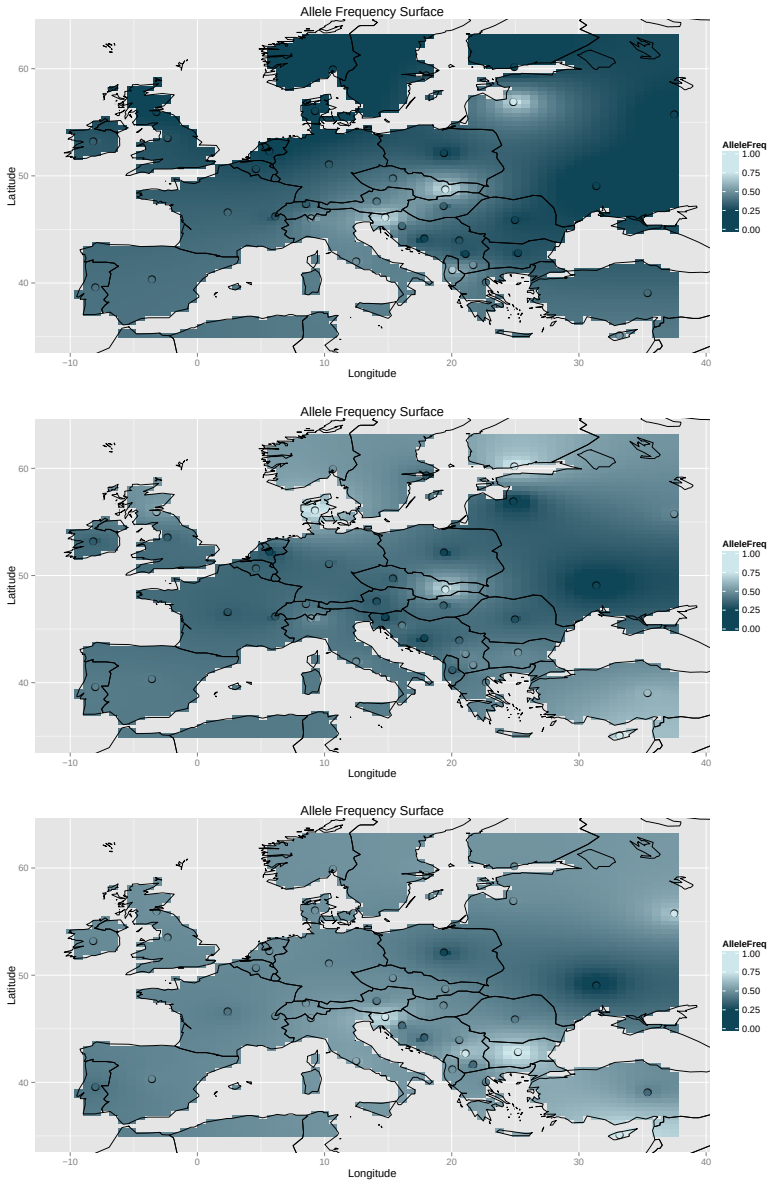


Figure 4.13: Additional estimated allele frequency surfaces for $\rho = 0.1$.

CHAPTER 5

Future Work

5.1 Landscape Genetics

5.1.1 Spatial Haplotypes

One easy modification we could make to the SNPscape model would be to look at haplotypes instead of SNPs. Although this may seem counterproductive, it has been shown in association studies that haplotype signatures can offer more definitive predictors than SNPs [AJX01, ASL07]. If we are dealing with phased data, we simply change from a Binomial model to a Multinomial model. The penalized loglikelihood becomes

$$f(\mathbf{p}) = \sum_i \ln\left(\frac{(2n)!}{[\prod_h n_{hi}!]}\prod_h p_{hi}^{n_{hi}}\right) - \rho\sum_h\sum_{i,j} \mathrm{w}_{ij}(p_{hi}-p_{hj})^2$$

where n is the total number of individuals measured, p_{hi} is the estimate of haplotype h in pixel i, w_{ij} is a weight between i and j, and n_{hi} is the number of h haplotypes in pixel i. This can be maximized via an MM algorithm and used the same way we use the SNPscape model.

Often, we do not have phase information. In this case, we can estimate haplotype frequencies using a penalized MM approach as done by Ayers [AL08]. In doing so, we would still "pixellate" the region of interest and bin the populations into their corresponding pixels. The penalized likelihood suggested by Ayers is

$$L(\mathbf{p}) = \sum_i \ln(r_i) - \lambda\sum_h f(p_h)\,,$$

where $r_i = \sum_{(k,l)\in H_i} p_k p_l$, H_i is the set of maternal-paternal haplotype pairs consistent with the observed genotypes of person i at the various markers, p_h is the frequency of haplotype

h, and

$$f(w_k) = \begin{cases} w_k & w_k \leq \delta \\ \delta & w_k \geq \delta \end{cases} .$$

The approach we would take would add another layer on top of this analysis. It would do this analysis for all pixels in the region with an additional penalty forcing haplotype frequencies of neighboring pixels together. This penalty allows us to estimate haplotype frequencies in regions with no data and also gives better estimates at locations with data by borrowing power from other locations with data. The penalized loglikelihood would be

$$L(\mathbf{p}) = \sum_j \left[\sum_i \ln(r_{ij}) - \lambda \sum_h f(p_{hj}) \right] - \rho \sum_h \sum_{j,k} \mathrm{w}_{jk} (p_{hj} - p_{hk})^2 ,$$

where the sum j is over all pixels in the region of interest and the sum over j, k in the penalty is non-zero only for a pixel's 8 nearest neighbors.

Another interesting problem when dealing with haplotypes is selection. In SNPscape, we choose the most informative SNPs based on a likelihood ratio test. We can do the same thing here, again replacing the binomial model with the more appropriate multinomial model, but we would also need to consider how many SNPs our haplotypes should contain. If we allow this number to vary this becomes an exponentially difficult problem.

5.1.2 Landscape Weighting

In the SNPscape model above, although the ocean tiles were weeded out of the analysis when placing individuals, it was considered an ordinary tile when propagating the allele frequency surface. In future work, where it is appropriate, we can modify different tiles to have different weights. For example, in the POPRES dataset we could reduce the water tile weights under the assumption that, for human populations, it takes longer to cross water than land. Alternatively, we could allow 0 crossing of water tiles but that leaves some of the islands in Europe with no data points. This improvement would allow for the analysis of datasets with interesting landscape features such as genetic barriers or bridges [MTS11, MGH04].

5.1.3 Individual vs. Group of Samples

Another interesting question we could consider is the placement of a group of samples we know came from the same location. For example, if there is a seizure of black market ivory in Africa, there is a good chance that they were all poached from the same location. So, by using the collective sample instead of each sample individually we may be able to pinpoint the location of the poaching more accurately.

Suppose that instead of having a group of individual samples, you have a group of samples which you believe came from the same population. Would this knowledge change where the samples are localized to? The answer is yes and here is a simple example to illustrate why this is the case. Suppose you have a population at point A whose allele at a given locus is all 1, another at B with $\frac{1}{2}$ allele 1 and $\frac{1}{2}$ allele 2, and another at C with all allele 2. Further suppose you have 2 samples one with allele 1 and the other with 2. Considering them as individual samples you would localize one to A and the other to C, whereas if you knew somehow that they came from the same population then they would have clearly been localized to B.

One nice result from our model is that we can easily get both probabilities. To get the probability that an entire group came from some population we again break it up by locus and multiply under the assumption of independence. For a given locus, if the group is unrelated, the probability that a group came from there is multinomial. Namely if x_h represents the number of h alleles at a locus, then the probability that the group came from a particular pixel is

$$P = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{j=1}^{k} p_j^{x_j} ,$$

where $\sum_{i=1}^{k} x_i = n$ and p_i is the estimated allele frequency at that pixel.

5.1.4 Sequence Data

In the previous sections, we had an implicit assumption that the SNP information we received was completely correct. This assumption may not always be a good one, so we can improve

upon the model by looking at sequence data instead. Under this model, we would have to estimate the probability an individual is homozygous for the minor allele, heterozygous, or homozygous for the major allele. The loglikelihood for an individual is

$$\ln(L) = \ln\left(p^2 \binom{n}{x} (1-\epsilon)^x \epsilon^{n-x} + 2pq \binom{n}{x} (\frac{1}{2})^n + q^2 \binom{n}{x} (1-\epsilon)^{n-x} \epsilon^x\right),$$

where n is the number of reads, x is the number of minor allele reads, ϵ is a small number, and p and q are the estimates for the probabilities that the individual has a minor allele and a major allele, respectively. This can be minorized by utilizing the concavity of ln. The minorization is

$$\begin{aligned}
\ln(L) &\geq \frac{p_m^2(1-\epsilon)^x\epsilon^{n-x}}{A_m} \ln(\frac{A_m p^2}{p_m^2}) + \frac{2p_m q_m (\frac{1}{2})^n}{A_m} \ln(\frac{A_m pq}{p_m q_m}) + \\
&\quad \frac{q_m^2(1-\epsilon)^{n-x}\epsilon^x}{A_m} \ln(\frac{A_m q^2}{q_m^2}) + \ln(\binom{n}{x}) \\
&= \frac{p_m^2(1-\epsilon)^x\epsilon^{n-x}}{A_m} \ln(p^2) + \frac{2p_m q_m (\frac{1}{2})^n}{A_m} \ln(pq) + \\
&\quad \frac{q_m^2(1-\epsilon)^{n-x}\epsilon^x}{A_m} \ln(q^2) + C \\
&= \frac{2\alpha_m}{A_m} \ln(p) + \frac{\beta_m}{A_m} \ln(pq) + \frac{2\gamma_m}{A_m} \ln(q) + C \\
&= \frac{2\alpha_m}{A_m} \ln(p) + \frac{\beta_m}{A_m} \ln(p(1-p)) + \frac{2\gamma_m}{A_m} \ln(1-p) + C \\
&= \frac{2\alpha_m + \beta_m}{A_m} \ln(p) + \frac{2\gamma_m + \beta_m}{A_m} \ln(1-p) + C,
\end{aligned}$$

where $\alpha_m = p_m^2(1-\epsilon)^x\epsilon^{n-x}$, $\beta_m = 2p_m q_m(\frac{1}{2})^n$, $\gamma_m = q_m^2(1-\epsilon)^{n-x}\epsilon^x$, $A_m = \alpha_m + \beta_m + \gamma_m$, and C is an irrelevant constant. Note that we can add in an ℓ_2 penalty for nearest neighbors and minorize it in the same way we did before. This gives a cubic equation that is very similar to the SNPscape model. In fact, the same arguments work in showing there is exactly one real solution in $(0, 1]$.

5.2 Landscape Measurements

Up to this point, we have limited ourselves to genetic data when there is no real need to do so. With a small modification to our underlying probability model, we can use this framework

to analyze many different types of data. With a Normal, or perhaps a Poisson, model we could look at flu data from Google [GMP08] or isotope data [MLT90]. Assuming the data is independent, we could even combine these different data to get a more accurate localization of unknown individuals. For example, using isotope data along with genetic data from ivory samples, we could localize poaching sites more accurately [MLT90, WJD08, WSC04]. More interestingly, it has been shown that the isotope ratios of Hydrogen and Oxygen in human hair are related to geography [EBC08]. Furthermore, it was shown that human travel can be tracked using these same ratios in hair [OW07]. Thus, in the future it would be interesting to combine isotope data with genetic data in the placement of humans on the map.

5.2.1 Gaussian distribution

For an isotope model we turn to the Normal probability distribution with a pseudo L1 penalty. This penalty is chosen because it is more lenient to widely varying differences, a problem we didn't need to worry about previously since frequencies are limited to $(0, 1)$. The penalty is taken to be $\sqrt{(\mu_i - \mu_j)^2 + \epsilon}$ for small ϵ so that the function is differentiable at 0. The Loglikelihood is

$$\ln L = \sum_k \left(-\frac{(x_k - \mu_k)^2}{2\sigma^2} - \frac{1}{2}\ln \sigma^2 - \frac{1}{2}\ln(2\pi) \right) - \sum_i \sum_{j \neq i} w_{ij} \sqrt{(\mu_i - \mu_j)^2 + \epsilon}$$

where x_k is the measurement at location k, σ^2 is the variance which is assumed to be the same for all locations, w_{ij} is a weight between i and j, and μ_k is the estimate at location k. To maximize this, we use the following majorization.

$$\sqrt{(x-y)^2 + \epsilon} \leq \sqrt{(x^m - y^m)^2 + \epsilon} + \frac{((x-y)^2 - (x^m - y^m)^2)}{2\sqrt{(x^m - y^m)^2 + \epsilon}}$$

This however does not separate parameters. To separate parameters, we can again turn to De Pierro focusing in on $(x-y)^2$.

$$(x-y)^2 \leq \frac{1}{2}[2x - x^m - y^m]^2 + \frac{1}{2}[2y - x^m - y^m]^2$$

Using the combined majorization we get the minorization

$$\begin{aligned}\ln L \geq & \sum_k \left(-\frac{(x_k - \mu_k)^2}{2\sigma^2} - \frac{1}{2}\ln \sigma^2 - \frac{1}{2}\ln(2\pi) \right) \\ & - \frac{1}{2}\sum_i \sum_{j \neq i} w_{ij} \left[\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon} \right. \\ & \left. + \frac{(\frac{1}{2}[2\mu_i - \mu_i^m - \mu_j^m]^2 + \frac{1}{2}[2\mu_j - \mu_i^m - \mu_j^m]^2 - (\mu_i^m - \mu_j^m)^2)}{2\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}} \right] \\ = & \, g(\mu | \mu^m).\end{aligned}$$

Taking derivatives yields

$$\frac{\partial g}{\partial \mu_i} = \frac{x_i - \mu_i}{\sigma^2} - \sum_{j \neq i} w_{ij} \frac{2\mu_i - \mu_i^m - \mu_j^m}{2\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}}.$$

Setting this equal to zero and solving gives our update for the estimated intensity μ.

$$\hat{\mu}_i = \frac{\frac{x_i}{\sigma^2} + \frac{1}{2}\sum_{j \neq i} \frac{w_{ij}(\mu_i^m + \mu_j^m)}{\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}}}{\frac{1}{\sigma^2} + \sum_{j \neq i} \frac{w_{ij}}{\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}}}$$

To get estimates for the variance, σ^2, we make the simplifying assumption that the variance is equal at all locations. Taking derivatives we then get

$$\frac{\partial \ln L}{\partial \sigma^2} = \sum_k \left(\frac{(x_k - \mu_k)^2}{2(\sigma^2)^2} - \frac{1}{2\sigma^2} \right).$$

Solving this equation we get the exact solution

$$\hat{\sigma}^2 = \frac{1}{\text{Total \# obs}} \sum_k (x_k - \mu_k)^2.$$

5.2.2 Poisson Model

For a Poisson model, useful for count data like Google trends, the loglikelihood is

$$\ln L = \sum_k n_k \ln(\mu_k) - \mu_k - \ln(n_k!) - \sum_i \sum_{j \neq i} w_{ij} \sqrt{(\mu_i - \mu_j)^2 + \epsilon}$$

where n_k is the count data for location k, w_{ij} is a weight between i and j, and μ_k is the estimate for location k. Using the combined majorization, as done above, we get the minorization

$$\begin{aligned}\ln L &\geq \sum_k \left(n_k \ln(\mu_k) - \mu_k - \ln(n_k!)\right) - \frac{1}{2}\sum_i \sum_{j\neq i} w_{ij} \left[\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}\right. \\ &\quad \left. + \frac{(\frac{1}{2}[2\mu_i - \mu_i^m - \mu_j^m]^2 + \frac{1}{2}[2\mu_j - \mu_i^m - \mu_j^m]^2 - (\mu_i^m - \mu_j^m)^2)}{2\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}}\right] \\ &= g(\mu|\mu^m).\end{aligned}$$

Taking derivatives yields

$$\frac{\partial g}{\partial \mu_i} = \frac{n_i}{\mu_i} - 1 - \sum_{j\neq i} w_{ij} \frac{2\mu_i - \mu_i^m - \mu_j^m}{2\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}}.$$

Setting this equal to zero and solving gives the quadratic equation

$$\left(\sum_{j\neq i} \frac{w_{ij}}{\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}}\right)\mu_i^2 + \left(1 - \sum_{j\neq i} \frac{w_{ij}(\mu_i^m + \mu_j^m)}{2\sqrt{(\mu_i^m - \mu_j^m)^2 + \epsilon}}\right)\mu_i - n_i = 0$$

and thus the updates. Using these updates, again pixellating first and penalizing only the nearest 8 neighbors, we can tackle a wide range of data types and combine different forms of data when localizing individuals.

5.2.3 Spatial-Temporal Measurements

In certain applications it may be beneficial to include a temporal element to the model. For example, when dealing with flu data from Google [GMP08], we would expect the number of people infected today in Los Angeles to not only be close to the number of infected today in Irvine but also to the number infected yesterday in Los Angeles. We could easily change the above model to give temporal measurements for each pixel and modify the penalty to include the nearest measurements in time as well as space. Alternatively, we could change from a univariate normal to a multivariate normal and include an autocorrelation model.

It has been shown that isotope ratios in human hair is related to geography and can be used to track human travel [OW07, EBC08]. This is due to the isotopes, which vary

with geography, getting incorporated into an individual's hair through drinking water. An interesting application to this model would be to localize an individual based on the differing isotope ratios in different positions, and thus different times, of their hair.

5.3 Random Multigraphs and Barrier Identification

An interesting problem in landscape genetics is the detection of barriers to genetic flow [DSE02, MGH04]. Identifying potential gene flow barriers is a major focus of landscape genetics research. While all landscape features affect gene flow, particular structures such as roads [RPS06], waterways [ASE06], or mountain ridges [FBC05] are potentially impenetrable barriers. These barriers can be tangible landscape features, but can also be intangible like cultural differences. The detection and incorporation of these barriers into models would be a good improvement over previous models. In previous work, we analyzed random multigraph networks and were able to detect significant overabundances of edges between two nodes. With enough data, we could also detect significant lacks of edges between nodes. Thus, if we could fit the genetic data of POPRES, for example, into a multigraph form the lacks of edges could imply a genetic barrier between the populations.

To get the data into a usable form, we need to convert our population frequencies into an integer number of connections between each pair of populations. One common measure of genetic distance between populations is the fixation index [Bro70]. A simple application of this index is

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}},$$

where π_{Between} represents the average number of pairwise differences between two individuals sampled from different subpopulations and π_{Within} represents the average from the same subpopulation. If we take instead the average number of pairwise similarities, instead of differences, between two individuals from different subpopulations, then we will have the data we need for the model.

Of course, there is a distance aspect to genetics as well. Thus the further apart two

populations are, the less we expect them to be similar. So we propose the following modification to the mean. Instead of $\lambda_{ij} = p_i p_j$ where p_i is the propensity of population i, we have $\lambda_{ij} = f(d_{ij})p_i p_j$ where $f(d_{ij})$ is some function of the distance between populations i and j. With these tools in hand, we could find genetic barriers by comparing the number of edges between two populations (nodes) to the mean number we expect under a Poisson distribution. One other interesting result from this analysis would be the population propensities. Although the meaning of these numbers is not immediately obvious, we would wager that they represent the "reach" of each population. That is, how far each population is able to spread their genetics. This would include several aspects such as the size of each population's empire and perhaps how much they traveled or tried to conquer other countries. Regardless, it is something that we would need to investigate further.

An interesting addition to this project would be to do the same analysis using a directed multigraph. In this case each population would have two propensities, one which relates to how much they spread their genetics and the other which relates to how much genetic material it takes in.

5.3.1 Bridge and Barrier Optimization

After discussing the identification of barriers to gene flow, the question that naturally arises next is how do we optimally reduce the affect of the barrier. For example, it has been shown that a California highway is a physical barrier to gene flow in carnivores [RPS06]. A common solution is to build animal bridges which are either overpasses or underpasses for animals to cross the highway safely. The number of these bridges needed and their locations is naturally an optimization problem.

In addition to the construction of bridges which assist in flow, we could also look into the construction of barriers to resist flow. This does not make too much sense in genetics, disregarding Monsanto, but in disease prevention it does. For example if an especially virulent strand of the flu were spreading, and it originated in Los Angeles, how could we best contain it? The answer is the optimal placement of barriers.

References

[AB02] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks." *Reviews of modern physics*, **74**(1):47, 2002.

[ABF08] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. "Mixed Membership Stochastic Blockmodels." *Journal of Machine Learning Research*, **9**:1981–2014, 2008.

[AJX01] Joshua Akey, Li Jin, Momiao Xiong, et al. "Haplotypes vs single marker linkage disequilibrium tests: what do we gain?" *European Journal of Human Genetics*, **9**(4):291–300, 2001.

[Aka74] H. Akaike. "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on*, **19**(6):716 – 723, dec 1974.

[AL08] Kristin L Ayers and Kenneth Lange. "Penalized estimation of haplotype frequencies." *Bioinformatics*, **24**(14):1596–1602, 2008.

[AL11] David H. Alexander and Kenneth Lange. "Enhancements to the ADMIXTURE Algorithm for Individual Ancestry Estimation." *BMC Bioinformatics*, **12**:246, 2011.

[ALG08] M Alfmova, T Lezhelko, V Golimbet, G Korovaltseva, O Lavrushkina, N Kolesina, L Frolova, A Muratova, L Abramova, and V Kaleda. "Investigation of association of the brain-derived neurotrophic factor (BDNF) and a serotonin receptor 2A (5-HTR2A) genes with voluntary and involuntary attention in schizophrenia." *Zh Nevrol Psikhiatr Im S S Korsakova*, **108**(4):62–9, 2008.

[ANL09] David H Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals." *Genome Research*, **19**(9):1655–1664, 2009.

[ASE06] M Antolin, L Savage, and R Eisen. "Landscape features influence genetic structure of black-tailed prairie dogs (Cynomys ludovicianus)." *Landscape Ecology*, **21**(6):867–875, 2006.

[ASL07] Kristin L Ayers, Chiara Sabatti, and Kenneth Lange. "A dictionary model for haplotyping, genotype calling, and association testing." *Genetic epidemiology*, **31**(7):672–683, 2007.

[AWP09] Sangtae Ahn, Richard T Wang, Christopher C Park, Andy Lin, Richard M Leahy, Kenneth Lange, and Desmond J Smith. "Directed mammalian gene regulatory networks using expression and comparative genomic hybridization microarray data from radiation hybrids." *PLoS computational biology*, **5**(6):e1000407, 2009.

[BA99] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks." *science*, **286**(5439):509–512, 1999.

[BRM09] Sebastian Bernhardsson, Luis Enrique Correa da Rocha, and Petter Minnhagen. "The meta book and size-dependent properties of written language." *New Journal of Physics*, **11**(12):123015, 2009.

[Bro70] AHD Brown. "The estimation of Wright's fixation index from genotypic frequencies." *Genetica*, **41**(1):399–406, 1970.

[CA09] Jennifer M Carbrey and Peter Agre. "Discovery of the aquaporins and development of the field." In *Aquaporins*, pp. 3–28. Springer, 2009.

[CES07] LS Chen, F EmmertStreib, and JD Storey. "Harnessing naturally randomized transcription to infer regulatory relationships among genes." *Genome Biol*, **8**(219), 2007.

[CHC06] Beth L Chen, David H Hall, and Dmitri B Chklovskii. "Wiring optimization can relate neuronal structure and function." *Proceedings of the National Academy of Sciences of the United States of America*, **103**(12):4723–4728, 2006.

[CL02] Fan Chung and Linyuan Lu. "The average distances in random graphs with given expected degrees." *Proceedings of the National Academy of Sciences*, **99**(25):15879–15882, 2002.

[CLR04] FS Collins, ES Lander, J Rogers, RH Waterston, and IHGS Conso. "Finishing the euchromatic sequence of the human genome." *Nature*, **431**(7011):931–945, 2004.

[CS05] T. Chan and J. Shen. *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*. Society for Industrial and Applied Mathematics, 2005.

[CS06] Ryan B Corcoran and Matthew P Scott. "Oxysterols stimulate Sonic hedgehog signal transduction and proliferation of medulloblastoma cells." *Proceedings of the National Academy of Sciences*, **103**(22):8408–8413, 2006.

[CS08] Math P Cuajungco and Mohammad A Samie. "The varitint–waddler mouse phenotypes and the TRPML3 ion channel mutation: cause and consequence." *Pflügers Archiv-European Journal of Physiology*, **457**(2):463–473, 2008.

[CTS00] I Cassens, R Tiedemann, F Suchentrunk, and GB Hartl. "Mitochondrial DNA variation in the European otter (Lutra lutra) and the use of spatial autocorrelation analysis in conservation." *Journal of Heredity*, **91**(1):31–35, Jan-Feb 2000.

[CZF06] M Carlson, B Zhang, Z Fang, P Mischel, S Horvath, and S F Nelson. "Gene Connectivity, Function, and Sequence Conservation: Predictions from Modular Yeast Co-expression Networks." *BMC Genomics*, **7**(7):40, 2006.

[DAS06] EJ Deeds, O Ashenberg, and EI Shakhnovich. "A simple physical model for scaling in protein-protein interaction networks." *Proc Natl Acad Sci U S A*, **103**(2):311–316, January 2006.

[DH77] Jan De Leeuw and Willem J Heiser. "Convergence of correction matrix algorithms for multidimensional scaling." *Geometric representations of relational data*, pp. 735–752, 1977.

[DH07] J Dong and S Horvath. "Understanding Network Concepts in Modules." *BMC Syst Biol*, **1**(1):24, 2007.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[DSE02] I Dupanloup, S Schneider, and L Excoffier. "A simulated annealing approach to define the genetic structure of populations." *Mol Ecol*, **11**(12):2571–81, Dec 2002.

[dup] "Wikipedia entry on Dupont." http://en.wikipedia.org/wiki/DuPont. Accessed: 14Dec2012.

[Dur06] Rick Durrett. *Random Graph Dynamics*. Cambridge University Press, New York, 2006.

[Dur07] Richard Durrett. *Random graph dynamics*, volume 20. Cambridge university press, 2007.

[DYK12] JA Dawson, S Ye, and C Kendziorski. "An empirical bayesian framework for discovering differential co-expression." *Bioinformatics*, [**68**](2):455–465, 2012.

[EBC08] J Ehleringer, G. Bowen, L. Chesson, A West, D Podlesak, and T Cerling. "Hydrogen and oxygen isotope ratios in human hair are related to geography." *Proceedings of the National Academy of Sciences*, **105**(8):2788–2793, 2008.

[ER59] Paul Erdős and Alfréd Rényi. "On random graphs." *Publicationes Mathematicae Debrecen*, **6**:290–297, 1959.

[ER60] Paul Erdos and A Rényi. "On the evolution of random graphs." *Publ. Math. Inst. Hungar. Acad. Sci*, **5**:17–61, 1960.

[ESB98] MB Eisen, PT Spellman, PO Brown, and D Botstein. "Cluster analysis and display of genome-wide expression patterns." *Proc Natl Acad Sci U S A*, **95**(25):14863–14868, December 1998.

[FBC05] C Funk, M Blouin, P Corn, BRYCE A. MAXELL, DAVID S. PILLIOD, STEPHEN AMISH, and FRED W. ALLENDORF. "Population structure of Columbia spotted frogs (Rana luteiventris) is strongly affected by the landscape." *Molecular Ecology*, **14**(2):483–496, 2005.

[FDC09] James H Fowler, Christopher T Dawes, and Nicholas A Christakis. "Model of genetic variation in human social networks." *Proceedings of the National Academy of Sciences*, **106**(6):1720–1724, 2009.

[FGA07] TF Fuller, A Ghazalpour, JE Aten, T Drake, AJ Lusis, and S Horvath. "Weighted gene coexpression network analysis strategies applied to mouse weight." *Mamm Genome*, **18**(6-7):463–472, 2007.

[Fis37] RA Fisher. "The wave of advance of advantageous genes." *Ann. Eugenics*, **7**:353–369, 1937.

[Fis00] R. A. Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, USA, 1 edition, April 2000.

[fre] "A community-curated database of well-known people, places, and things." `http://www.freebase.com/`. Accessed: 15May2012.

[Fue10] A de la Fuente. "From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases." *Trends Genet*, **26**(7):326–333, 2010.

[GBH03] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. "The international HapMap project." *Nature*, **426**(6968):789–796, 2003.

[GCV07] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. "The human disease network." *Proceedings of the National Academy of Sciences*, **104**(21):8685–8690, 2007.

[GDZ06] A Ghazalpour, S Doss, B Zhang, C Plaisier, S Wang, E Schadt, A Thomas, T Drake, A Lusis, and S Horvath. "Integrating genetics and network analysis to characterize genes related to mouse weight." *PloS Genetics*, **2**(8), 2006.

[Gen] Johns Hopkins University McKusick-Nathans Institute of Genetic Medicine. "Online Mendelian Inheritance in Man, OMIM®.".

[GGC06] Peter S. Gargalovic, Nima M. Gharavi, Michael J. Clark, Joanne Pagnon, Wen-Pin Yang, Aiqing He, Amy Truong, Tamar Baruch-Oren, Judith A. Berliner, Todd G. Kirchgessner, and Aldons J. Lusis. "The Unfolded Protein Response Is an Important Regulator of Inflammatory Genes in Endothelial Cells." *Arterioscler Thromb Vasc Biol*, **26**(11):2490–2496, 2006.

[GH75] Stephen J Goss and Henry Harris. "New method for mapping genes in human chromosomes." 1975.

[GMP08] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. "Detecting influenza epidemics using search engine query data." *Nature*, **457**(7232):1012–1014, 2008.

[HD08] S Horvath and J Dong. "Geometric interpretation of Gene Co-expression Network Analysis." *PloS Comput Biol*, **4**:8, 2008.

[HF95] David I Holmes and Richard S Forsyth. "The Federalist revisited: New directions in authorship attribution." *Literary and Linguistic Computing*, **10**(2):111–127, 1995.

[HL81] Paul W Holland and Samuel Leinhardt. "An exponential family of probability distributions for directed graphs." *Journal of the american Statistical association*, **76**(373):33–50, 1981.

[HL00] David R Hunter and Kenneth Lange. "Quantile regression via an MM algorithm." *Journal of Computational and Graphical Statistics*, **9**(1):60–77, 2000.

[HL04] David R Hunter and Kenneth Lange. "A tutorial on MM algorithms." *The American Statistician*, **58**(1):30–37, 2004.

[HLH07] Y Huang, H Li, H Hu, X Yan, MS Waterman, H Huang, and XJ Zhou. "Systematic discovery of functional modules and context-specific functional annotation of human genome." *Bioinformatics*, **23**(13):i222–229, 2007.

[HLL83] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps." *Social networks*, **5**(2):109–137, 1983.

[Hof07] Peter D Hoff. "Modeling homophily and stochastic equivalence in symmetric relational data." *arXiv preprint arXiv:0711.1146*, 2007.

[Hor11] Steve Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer, 1 edition, May 2011.

[HRH02] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. "Latent space approaches to social network analysis." *Journal of the american Statistical association*, **97**(460):1090–1098, 2002.

[HSA05] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." *Nucleic acids research*, **33**(suppl 1):D514–D517, 2005.

[HW08] Jake M Hofman and Chris H Wiggins. "Bayesian approach to network modularity." *Phys Rev Lett*, **100**(25):258701, Jun 2008.

[HZC06] S Horvath, B Zhang, M Carlson, K Lu, S Zhu, R Felciano, M Laurance, W Zhao, Q Shu, Y Lee, A Scheck, L Liau, H Wu, D Geschwind, P Febbo, H Kornblum, T Cloughesy, S Nelson, and P Mischel. "Analysis of oncogenic signaling networks in Glioblastoma identifies ASPM as a novel molecular target." *Proc Natl Acad Sci USA*, **103**(46):17402–17407, 2006.

[KCW08] MP Keller, YJ Choi, P Wang, D Belt Davis, ME Rabaglia, AT Oler, DS Stapleton, C Argmann, KL Schueler, S Edwards, HA Steinberg, Elias Chaibub Neto, Robert Kleinhanz, Scott Turner, Marc K Hellerstein, EE Schadt, BS Yandell,

C Kendziorski, and AD Attie. "A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility." *Genome Res*, **18**(5):706–716, 2008.

[KM90] Suwon Kim and Richard M Myers. "Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes." 1990.

[KPP37] A. Kolmogorov, I. Petrovskii, and N. Piscunov. "A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem." *Byul. Moskovskogo Gos. Univ.*, **1**(6):1–25, 1937.

[KR90] L Kaufman and PJ Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley and Sons, Inc, 1990.

[KRL00] P. L. Krapivsky, S. Redner, and F. Leyvraz. "Connectivity of Growing Random Networks." *Phys. Rev. Lett.*, **85**:4629–4632, Nov 2000.

[KTG06] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. "Learning Systems of Concepts with an Infinite Relational Model." In *AAAI*, pp. 381–388. AAAI Press, 2006.

[Lan90] Kenneth Lange. "Convergence of EM image reconstruction algorithms with Gibbs smoothing." *Medical Imaging, IEEE Transactions on*, **9**(4):439–446, 1990.

[Lan99] Kenneth Lange. *Numerical Analysis for Statisticians*. Springer-Verlag, New York, 1999.

[Lan04] K Lange. *Optimization*. Springer-Verlag, New York, 2004.

[Lan10] K. Lange. *Numerical analysis for statisticians*. Springer, 2010.

[Lan12] K. Lange. *Numerical Analysis for Statisticians*. Statistics and Computing. Springer London, Limited, 2012.

[LH07] P Langfelder and S Horvath. "Eigengene networks for studying the relationships between co-expression modules." *BMC Syst Biol*, **1**(1):54, 2007.

[LH08] P Langfelder and S Horvath. "WGCNA: an R package for weighted correlation network analysis." *BMC Bioinformatics*, **9**(1):559, 2008.

[LHY00] Kenneth Lange, David Hunter, and Ilsoon Yang. "Optimization Transfer Using Surrogate Objective Functions." *Journal of Computational and Graphical Statistics*, **9**(1), 2000.

[LLB01] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. "Initial sequencing and analysis of the human genome." *Nature*, **409**(6822):860–921, 2001.

[LLN08] Oscar Lao, Timothy T Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A Bindoff, David Comas, et al. "Correlation between genetic and geographic structure in Europe." *Current Biology*, **18**(16):1241–1248, 2008.

[Lux07] Ulrike von Luxburg. "A tutorial on spectral clustering." *Statistics and Computing*, **17**(4):395–416, 2007.

[LZH07] P Langfelder, B Zhang, and S Horvath. "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R." *Bioinformatics*, **24**(5):719–20, 2007.

[MGH04] F Manni, E Guerard, and E Heyer. "Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected using Monmonier's algorithm." *Human Biology*, **76**:173–190, 2004.

[MHK05] Steven Maere, Karel Heymans, and Martin Kuiper. "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks." *Bioinformatics*, **21**(16):3448–3449, 2005.

[MLT90] NJ Van der Merwe, JA Lee-Thorp, JF Thackeray, A Hall-Martin, FJ Kruger, H Coetzee, RHV Bell, and M Lindeque. "Source-area determination of elephant ivory by isotopic analysis." *Nature*, **346**(6286):744–746, 1990.

[MMS08] Christopher A Maxwell, Víctor Moreno, Xavier Solé, Laia Gómez, Pilar Hernández, Ander Urruticoechea, Miguel Angel Pujana, et al. "Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment." *Mol Cancer*, **7**(4), 2008.

[MS02] Sergei Maslov and Kim Sneppen. "Specificity and stability in topology of protein networks." *Science*, **296**(5569):910–913, 2002.

[MTS11] AGK Mirams, EA Treml, JL Shields, L Liggins, and C Riginos. "Vicariance and dispersal across an intermittent barrier: population genetic structure of marine animals across the Torres Strait land bridge." *Coral reefs*, **30**(4):937–949, 2011.

[MW83] William McColly and Dennis Weier. "Literary attribution and likelihood-ratio tests: the case of the middle English PEARL-poems." *Computers and the Humanities*, **17**(2):65–75, 1983.

[MW84] Frederick Mosteller and David Wallace. "Applied Bayesian and classical inference the case of the Federalist papers." 1984.

[NBK08] Matthew R Nelson, Katarzyna Bryc, Karen S King, Amit Indap, Adam R Boyko, John Novembre, Linda P Briley, Yuka Maruyama, Dawn M Waterworth, Gérard Waeber, et al. "The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research." *The American Journal of Human Genetics*, **83**(3):347–358, 2008.

[New06] Mark EJ Newman. "Modularity and community structure in networks." *Proceedings of the National Academy of Sciences*, **103**(23):8577–8582, 2006.

[NIHa] "U.S. National Institute of Healths informational page on Craniofacial-deafness-hand syndrome." `http://ghr.nlm.nih.gov/condition/craniofacial-deafness-hand-syndrome`. Accessed: 14Dec2012.

[NIHb] "U.S. National Institute of Healths informational page on Usher syndrome." `http://www.nidcd.nih.gov/health/hearing/pages/usher.aspx`. Accessed: 14Dec2012.

[NIHc] "U.S. National Institute of Healths informational page on Waardenburg syndrome." http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002401/. Accessed: 14Dec2012.

[NJB08] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. "Genes mirror geography within Europe." *Nature*, **456**(7218):98–101, Nov 2008.

[NL07] M E J Newman and E A Leicht. "Mixture models and exploratory analysis in networks." *Proc Natl Acad Sci U S A*, **104**(23):9564–9, Jun 2007.

[NS01] Krzysztof Nowicki and Tom A B Snijders. "Estimation and prediction for stochastic blockstructures." *Journal of the American Statistical Association*, **96**(455):1077–1087, 2001.

[NSW01] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. "Random graphs with arbitrary degree distributions and their applications." *Physical Review E*, **64**(2):026118, 2001.

[OHG06] Michael C Oldham, Steve Horvath, and Daniel H Geschwind. "Conservation and evolution of gene coexpression networks in human and chimpanzee brains." *Proceedings of the National Academy of Sciences*, **103**(47):17973–17978, 2006.

[OKI08] MC Oldham, G Konopka, K Iwamoto, P Langfelder, T Kato, S Horvath, and DH Geschwind. "Functional organization of the transcriptome in human brain." *Nat Neurosci*, **11**(11):1271–1282, October 2008.

[OR00] James M Ortega and Werner C Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Number 30. Siam, 2000.

[OW07] Diane M O'Brien and Matthew J Wooller. "Tracking human travel using stable oxygen and hydrogen isotope analyses of hair and urine." *Rapid Communications in Mass Spectrometry*, **21**(15):2422–2430, 2007.

[PAB08] Christopher C Park, Sangtae Ahn, Joshua S Bloom, Andy Lin, Richard T Wang, Tongtong Wu, Aswin Sekar, Arshad H Khan, Christine J Farr, Aldons J Lusis, et al. "Fine mapping of regulatory loci for mammalian gene expression using radiation hybrids." *Nature genetics*, **40**(4):421–429, 2008.

[PGK09] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. "Human protein reference database2009 update." *Nucleic acids research*, **37**(suppl 1):D767–D772, 2009.

[QSC12] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." *BMC genomics*, **13**(1):341, 2012.

[RAS10] John MO Ranola, Sangtae Ahn, Mary Sehl, Desmond J Smith, and Kenneth Lange. "A Poisson model for random multigraphs." *Bioinformatics*, **26**(16):2004–2011, 2010.

[RLL13] John Michael Ranola, Peter Langfelder, Kenneth Lange, and Steve Horvath. "Cluster and propensity based approximation of a network." *BMC systems biology*, **7**(1):21, 2013.

[RLW03] Noah A Rosenberg, Lei M Li, Ryk Ward, and Jonathan K Pritchard. "Informativeness of genetic markers for inference of ancestry." *The American Journal of Human Genetics*, **73**(6):1402–1422, 2003.

[Rog63] F Rogers. "Medical subject headings." *Bulletin of the Medical Library Association*, **51**(1):114–116, 1963.

[RPS06] S Riley, J Pollinger, R Sauvajot, E York, C Bromley, T Fuller, and R Wayne. "A southern California freeway is a physical and social barrier to gene flow in carnivores." *Mol Ecol*, **15**(7):1733–41, Jun 2006.

[SAK08] J Sinkkonen, J Aukia, and S Kaski. "Component Models for Large Networks." *arXiv e-prints*, **arXiv:0803.1628**, 2008.

[Sch78] Gideon Schwarz. "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2):pp. 461–464, 1978.

[Sch07] Satu Elisa Schaeffer. "Graph clustering." *Computer Science Review*, **1**(1):27 – 64, 2007.

[SDR06] S Steinberg, G Dodt, G Raymond, N Braveman, A Moser, and H Moser. "Peroxisome Biogenesis Disorders." *Biochemica et Biophysica Acta - Molecular Cell Research*, **1763**(12):1733, 2006.

[Sha83] B Shaanan. "Structure of Human Oxyhaemoglobin at 2.1 A Resolution." *Journal of Molecular Biology*, **171**(1):31–59, 1983.

[SO78] Robert R Sokal and Neal L Oden. "Spatial autocorrelation in biology: 1. Methodology." *Biological Journal of the Linnean Society*, **10**(2):199–228, 1978.

[SSK03] J M Stuart, E Segal, D Koller, and S K Kim. "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules." *Science*, **302**(5643):249–255, 2003.

[Ste05] U et al. Stelzl. "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome." *Cell*, **122**(6):957–68, 2005.

[Str01] Steven H Strogatz. "Exploring complex networks." *Nature*, **410**(6825):268–276, 2001.

[TMF09] Tracy Tucker, Marco Marra, and Jan M Friedman. "Massively parallel sequencing: the next big thing in genetic medicine." *The American Journal of Human Genetics*, **85**(2):142–154, 2009.

[Tob70] Waldo R Tobler. "A computer movie simulating urban growth in the Detroit region." *Economic geography*, **46**:234–240, 1970.

[UL01] Jean-Baptiste Hiriart Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer, 2001.

[Wal01] Bruce Walsh. "Quantitative genetics in the age of genomics." *Theoretical population biology*, **59**(3):175–184, 2001.

[Wat09] Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press, 2009.

[WJD08] Samuel K Wasser, William Joseph Clark, Ofir Drori, Emily Stephen Kisamo, Celia Mailand, Benezeth Mutayoba, and Matthew Stephens. "Combating the illegal trade in African elephant ivory with DNA forensics." *Conservation Biology*, **22**(4):1065–1071, 2008.

[WL10] T.T. Wu and K. Lange. "The MM alternative to EM." *Statistical Science*, **25**(4):492–505, 2010.

[WS98] DJ Watts and SH Strogatz. "Collective dynamics of 'small-world' networks." *Nature*, **393**(6684):440–2, 1998.

[WSC04] Samuel K Wasser, Andrew M Shedlock, Kenine Comstock, Elaine A Ostrander, Benezeth Mutayoba, and Matthew Stephens. "Assigning African elephant DNA to geographic region of origin: applications to the ivory trade." *Proceedings of the National Academy of Sciences of the United States of America*, **101**(41):14847–14852, 2004.

[WST86] John G White, Eileen Southgate, J Nichol Thomson, and Sydney Brenner. "The structure of the nervous system of the nematode Caenorhabditis elegans." *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **314**(1165):1–340, 1986.

[WW87] Yuchung J Wang and George Y Wong. "Stochastic blockmodels for directed graphs." *Journal of the American Statistical Association*, **82**(397):8–19, 1987.

[XL10] Ramon Xulvi-Brunet and Hongzhe Li. "Co-expression networks: graph properties and topological comparisons." *Bioinformatics*, **26**(2):205–14, Jan 2010.

[YL11] J Yin and H Li. "A sparse conditional gaussian graphical model for analysis of genetical genomics data." *Annals of Applied Statistics*, **5**(4):2630–2650, 2011.

[YNE12] Wen-Yun Yang, John Novembre, Eleazar Eskin, and Eran Halperin. "A model-based approach for analysis of spatial structure in genetic data." *Nature genetics*, **44**(6):725–731, 2012.

[ZAL11] Hua Zhou, David Alexander, and Kenneth Lange. "A quasi-Newton acceleration for high-dimensional optimization algorithms." *Statistics and computing*, **21**(2):261–273, 2011.

[ZB07] Dengyong Zhou and Christopher J. C. Burges. "Spectral clustering and transductive learning with multiple views." In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pp. 1159–1166, New York, NY, USA, 2007. ACM.

[ZH05] B Zhang and S Horvath. "A general framework for weighted gene co-expression network analysis." *Statistical Applications in Genetics and Molecular Biology*, **4**:17, 2005.

[Zip32] G Zipf. "Selective Studies and the Principle of Relative Frequency in Language (Cambridge, Mass, 1932)." *Human Behavior and the Principle of Least-Effort (Cambridge, Mass*, 1932.