Project Description

The proposed project concerns problems related to the classification of three-dimensional manifolds and applications of the relevant methods to the analysis of large data sets. The pure and applied aspects of this project will complement and enhance each other, with each side suggesting open problems and new approaches for the other side.

1. Project Outline

Analyze the structure of the different Heegaard splittings admitted by any given three-dimensional manifold. While Heegaard splittings were defined over a hundred years ago, the tools to seriously analyze them have only been developed in the last three decades. A number of recent breakthroughs in three-dimensional hyperbolic geometry have both suggested new approaches to understanding Heegaard splittings and have demonstrated that Heegaard splittings are central to understanding three-dimensional geometry and topology. The PI proposes to expand and refine the existing techniques in light of the new directions suggested by hyperbolic geometry.

Generalize known results about Heegaard splittings to immersed surfaces and finite covers. A finite covering map sends every embedded surface to an immersed surface. Now that the tools that were developed to analyze (embedded) Heegaard surfaces are better understood, they can be generalized and applied to immersed surfaces. The PI proposes to translate these techniques as far as possible and apply the resulting methods to study finite covers.

Apply tools and ideas from low-dimensional topology to the analysis of large data sets. In the last few years, topology has been applied to the analysis of large data sets in a number of different ways. The PI has recently used thin position, the main technique underlying the first two components of this proposal, to develop an effective algorithm for decomposing data sets into clusters. The algorithm suggests a strong analogy between geometric topology and high dimensional data analysis and the PI proposes to further develop this analogy in order to apply more techniques from low dimensional/geometric topology to data analysis.

The theme that unites the three components of this project is the notion of *thin position*. This technique was first defined in order to study knots, but it has been generalized and applied in a number of different areas. Each component of this proposal will apply thin position to a different set of problems, and success in any one component will provide valuable insight into how thin position can be applied in the other components.

2. The structure of Heegaard splittings

Heegaard splittings were first defined in order to construct examples of three-dimensional manifolds related to a number of conjectures in the series of papers in which Poincaré initiated the field of three-dimensional topology. Since then, Heegaard splittings have become a central and fundamental part of the field, with connections to hyperbolic geometry, algebraic topology and to knot theory.

1. **Definition.** A *handlebody* is a three-dimensional manifold homeomorphic to a regular neighborhood of an embedded graph in three-dimensional space. A *Heegaard splitting* of a three-dimensional manifold M is a decomposition $M = H^- \cup_\Sigma H^+$ where H^- and H^+ are handlebodies that are glued together along their boundaries to form M. The surface $\Sigma \subset M$ is the image of the two boundary surfaces and is called a *Heegaard surface*.

There is a similar definition for Heegaard splittings of three-dimensional manifolds with boundary, but we will not include it here. The structure of the set of Heegaard splittings for a manifold M has two components:

The first component of the structure is the relationships between different Heegaard splittings of M. Reidemeister [38] and Singer [44] showed that for any two Heegaard splittings of M, one can turn the first splitting into the other by repeating a construction called *stabilization* and its inverse. Stabilization consists of adding a trivial handle to a Heegaard surface, producing a higher genus Heegaard surface.

2. **Definition.** The *Heegaard tree* for M is the graph in which each vertex represents an isotopy class of Heegaard splittings for M and each edge represents an instance of the stabilization construction.

Recent work by Hass-Thompson-Thurston [13], Bachman [2] and the PI [20] has demonstrated that Heegaard trees can have very different structures depending on the topology and geometry of M. These constructions are closely related to the hyperbolic geometric picture suggested by Biringer-Souto [4] and Namazi-Souto [37]. Relatively little is known in general about how complex a Heegaard tree can be, and how complex an arbitrary Heegaard tree is. However, techniques for classifying and analyzing Heegaard splittings have been steadily improving, particularly due to the above-mentioned connections to hyperbolic geometry.

The second component of the structure is the group of symmetries of each Heegaard splitting.

3. **Definition.** The *mapping class group* $Mod(M, \Sigma)$ of a Heegaard splitting $M = H^- \cup_\Sigma H^+$ is the group of self-homeomorphisms $\phi : M \to M$ such that $\phi(\Sigma) = \Sigma$, modulo isotopies of M that preserve Σ setwise.

Mapping class groups of Heegaard splittings have only recently been studied in their full generality. However, specific instances have been studied in a number of situations. The most intriguing instance began with a proof by Goeritz [11] that the mapping class group of the genus-two Heegaard splitting of the three-sphere is finitely generated with a very simple generating set. This was motivated by a number of applications to unknotting tunnels of knots. Two papers were later published claiming to generalize this to higher genus Heegaard splittings of the three-sphere, but fatal flaws have since been pointed out in both.

Many of the recently discovered techniques for understanding how different Heegaard splittings are related to each other can be applied to understanding mapping class groups and the PI has used these techniques to generate a number of interesting examples [23, 25, 26, 28]. The PI and Hyam Rubinstein have characterised mapping class elements corresponding to reducible automorphisms of the Heegaard surface [30]. As with the relations between different Heegaard splittings, these results suggest strong connections to the hyperbolic geometric picture in the work of Biringer-Souto [4] and Namazi-Souto [37].

Proposed Research Plans: In the Heegaard tree of a manifold M, each vertex can be labeled with the genus of the corresponding Heegaard surface. If we think of this as a height function on the graph then the height will change by one along each edge and there will be exactly one edge going up from each vertex. There are many different trees with height functions of this form, most of which will never appear as a Heegaard tree. For example, Rubinstein-Scharlemann [39] proved a bound on the lengths of the legs of any Heegaard tree (in terms of the genera of the feet), which was recently improved by the PI [22]. However, within these bounds relatively little is known about what graphs can appear as Heegaard trees.

4. **Motivating Problem.** Determine which trees can be realized as the Heegaard tree of a three-dimensional manifold.

We do not propose to solve this problem in the near future, but there are a number of more tractable problems that will shed light on it. The constructions used by Hass-Thompson-Thurston [13], Bachman [2] and the PI [20] to produce manifolds with complex Heegaard trees use techniques that involve Hempel distance [16]: an integer d assigned to a Heegaard surface Σ that measures how "complicated" the gluing map between the two handlebodies H^- and H^+ is. All three constructions generalize the following Theorem of Scharlemann and Tomova:

5. **Theorem** (Scharlemann-Tomova [43])**.** *If $H^- \cup_\Sigma H^+ = M$ is a Heegaard splitting with Hempel distance d then the Heegaard tree for M has at most one vertex at each height corresponding to a genus less than $\frac{d}{2}$.*

Hempel [16] showed that there are Heegaard splittings of any fixed genus $g \geq 2$ with arbitrarily high distance, so Theorem 5 implies that there are manifolds whose Heegaard tree has a solitary bottom leg of arbitrary length. However, the Theorem does not imply anything about the number of Heegaard splittings of genus greater than $\frac{d}{2}$ and there are similar constraints on the techniques used in all the previously mentioned constructions [2, 13, 20].

1. **Problem.** For every genus $g \geq 2$, is there a constant K_g such that if $H^- \cup_\Sigma H^+ = M$ is a Heegaard splitting with Hempel distance $d > K_g$ then the Heegaard tree for M has at most one vertex at EVERY height?

An affirmative answer to this question could most likely be generalized to strengthen the other known constructions (such as [2, 13, 20]). A negative answer would produce manifolds with very interesting Heegaard trees.

In order to address Problem 1, the PI proposes a more careful analysis of the connection between hyperbolic geometry and Heegaard surfaces. Hass-Thompson-Thurston's generalization of Theorem 5 replaces the notion of high Hempel distance with the following construction, which has also been studied by Namazi-Souto [37]:

Consider a hyperbolic surface bundle M and let $\hat{M}$ be the infinite cyclic cover defined by the bundle structure. The manifold $\hat{M}$ is a surface cross an interval, but it inherits a hyperbolic metric lifted from M. Cut out a compact subset of $\hat{M}$ consisting of a large number of fundamental domains of the cover. Namazi-Souto have shown that one can glue in handlebodies on both ends and extend the hyperbolic metric to an ϵ-pinched hyperbolic metric on the entire closed manifold where ϵ decreases as the number of fundamental domains in $\hat{M}$ increases.

The Heegaard splittings constructed this way are topologically essentially the same as the splittings constructed by Hempel with arbitrarily high distance [16].

Hass-Thompson-Thurston's proof draws a strong analogy between the geometric and topological pictures. However, both approaches rely on rather blunt instruments and lead to the same issue identified in Problem 1. On the other hand, Namazi-Souto have refined the surface bundle construction using a structure from Minsky's proof of the ending lamination conjecture [36], which the PI has begun adapting to the study of Heegaard surfaces [24].

6. **Definition.** A *model block* B is a manifold of the form $F \times [0,1]$ where F is either a four-punctured sphere or a once-punctured torus. In addition to the product structure, each model block is marked with one essential loop in each of $F \times \{0\}$ and $F \times \{1\}$, such that the projections are distinct but intersect minimally (either once or twice depending on F). These loops and the loop(s) $\partial F \times \{\frac{1}{2}\}$ cut the boundary of B into either two or four pairs of pants (three-holed spheres), and we form a *model structure* by gluing model blocks together along these pairs of pants.

The model structure defines a manifold in which the boundary loop of the pairs of pants form a link. Ends of hyperbolic manifolds can be approximated by gluing together infinitely many blocks [36], or finite volume manifolds can be constructed with finitely many blocks [24]. Namazi-Souto have proposed using this construction to construct manifolds with precisely determined geometric structures.

On the other hand, the PI has shown that compact model structures are in many ways analogous to triangulations, particularly the *layered triangulations* defined by Jaco-Rubinstein-Tillman [17]. However, because model structures are more closely tied to hyperbolic geometry, one would expect them to be more compatible with the topology and to avoid many of the limitations of three-dimensional triangulations.

Triangulations have been used extensively to study Heegaard splittings via normal surface theory [40, 45]. The PI has used techniques from almost normal surface theory to generalize Theorem 5 while translating the sweep-out methods in the topological proof into the piecewise-linear category [26, 23]. The PI is currently working with Bus Jaco to apply ideas from the proof of Theorem 5 to understanding normal surfaces in layered triangulations. These projects suggest that a version of normal surface theory for layered models should be both feasible and highly useful.

Rubinstein and Stockings' proofs [40, 45] that Heegaard surfaces can be put into almost normal form suggest immediate generalizations to model structures. The normal surface theory for a model structure constructed from a Heegaard splitting would likely lead to a generalization of Theorem 5 without the genus limitation.

1. **Objective.** The PI will work with Alex Zupan (an NSF postdoc at UT Austin) to develop normal surface theory for layered models and apply the resulting theory to construct manifolds whose Heegaard trees can be completely determined.

There are also much more fundamental questions about model structures that can be studied by analogy to triangulations. For example, are there finitely many moves that can turn a given model into any other model? What topological conditions on the link are equivalent to a model being minimal?

2. **Objective.** The PI will supervise Pengcheng Xu (a third year PhD student at OSU) in the development of ideas such as the analog of Pachner moves for model structures.

Problem 1 has to do with the length of the legs of the Heegaard tree. One might also wonder about the number of legs. Li [33] showed that every hyperbolic 3-manifold has finitely many Heegaard splittings of any given genus, i.e. that each horizontal row of its Heegaard tree will have finitely many vertices. On the other hand, Lustig-Moriah constructed hyperbolic manifolds with arbitrarily many minimal genus Heegaard splittings, showing that the number of vertices in the bottom row can be as large as one likes. In their construction, the minimal genus grows as the number of minimal genus Heegaard splittings grows, which begs the following question:

2. **Problem.** For every genus $g \geq 2$, is there a constant K_g such that if M is hyperbolic then M admits at most K_g isotopy classes of genus g Heegaard splittings?

Many three-dimensional manifolds with multiple distinct Heegaard splittings are can be constructed by gluing a number of three-dimensional pieces by high distance maps between their boundaries. This approach to studying hyperbolic manifolds is justified by Biringer-Souto's recent result that all compact, orientable, hyperbolic three-dimensional manifolds with bounded rank and injectivity radius can be constructed by gluing together pieces of a finite number of topological types along their boundaries [4]. As in Objective 1, these gluing constructions can be made more precise by using model structures. However, there is also a matter of understanding the choices of pieces and the pairing of their boundaries.

Tao Li has shown that for sufficiently high distance gluings, the (low genus) Heegaard splittings of resulting manifold are determined by the pieces in a natural way [33]. The PI has developed a program to strengthen these results to determine exactly what the Heegaard tree (below some fixed genus depending on the distance) looks like for such a manifold [27]. (In particular, the goal is to determine when distinct surfaces remain distinct up to isotopy.) The proof will combine Li's approach [33] with techniques developed by the PI to distinguish isotopy classes of Heegaard surfaces [20, 23, 26].

3. **Objective.** The PI will complete his program to determine the Heegaard tree (below a fixed genus) of any manifold that results from gluing together a finite number of pieces along high distance maps.

The completion of Objective 3 will reduce Problem 2 to a matter of understanding the Heegaard splittings of the pieces in these constructions, which are topologically simpler. A successful completion of both Objectives 1 and 3 should suggest a large family of constructions producing manifolds with interesting Heegaard trees that can be completely determined (i.e. without the genus limitation).

The simplest instance of such a gluing construction is Dehn filling, in which one glues a solid torus to a second, more complex piece. A number of open problems about Dehn surgery, such as the Cabling Conjecture and the Berge Conjecture (which we will not describe here) have to do with surfaces in the filled and unfilled manifolds. These require a much more precise control of different gluing maps, as opposed to the essentially asymptotic approaches (i.e. high distance gluing) described above. The following problem, first asked by Yoav Moriah, is closely related to both the Berge and Cabling conjectures:

3. **Problem.** Is there a manifold M with one torus boundary component T such that for two distinct slopes $\sigma, \tau \subset T$, Dehn filling T along each of σ, τ produces a 3-manifold whose Heegaard genus is at least two less than that of M?

A negative answer would imply that every knot in the 3-sphere with a Lens space surgery has tunnel number one (which would be positive progress on the Berge conjecture). An affirmative answer would show that if the Berge conjecture is true, it is a phenomenon very specific to the three-sphere.

The PI has been studying related problems with Ryan Blair, Marion Campisi, Scott Taylor and Maggy Tomova on a project supported by the AIM SQuaREs program. The first preprint resulting from this project [5] shows that if a link complement has a bridge surface with high distance then no non-trivial Dehn filling will produce manifolds with Heegaard genus less than that of the knot complement. This gives a very precise condition on the Dehn filling slope, but still requires a "high distance" condition on the bridge surface. We are currently developing an approach to studying low distance bridge surfaces, that should allow us to get very precise control of both the bridge distance and the Dehn filling slope.

4. **Objective.** The PI, with Ryan Blair, Marion Campisi, Scott Taylor and Maggy Tomova, will characterize low-distance bridge surfaces and apply this characterization to Dehn filling results.

As noted above, the second aspect of the structure of Heegaard splittings is their mapping class groups. In addition to defining very interesting subgroups of surface mapping class groups, the PI has shown [23, 25, 26, 28] that the mapping class groups of many Heegaard splittings are fundamentally representative of the geometric/topological structure of the ambient manifold. One of the most vexing problems related to mapping class groups of Heegaard splittings is to understand the mapping class groups of higher genus splittings of the three-sphere.

7. **Motivating Problem** (The Goeritz Problem)**.** Is the mapping class group of every genus $g > 2$ Heegaard splitting finitely generated?

This problem is far from tractable with current technology. Because such Heegaard splittings are highly stabilized, the resulting mapping class groups are very large and unwieldy. The PI and Hyam Rubinstein have proposed studying mapping class groups of unstabilized Heegaard splittings as a preliminary step towards the Goeritz problem [30]. In particular, we proved a characterization of mapping classes whose restrictions to the Heegaard surface are reducible surface automorphisms, under the assumption that the Heegaard splitting satisfies a condition called *strongly irreducible*. Such mapping classes fall into a number of categories determined by the topology of the ambient manifold and the PI has constructed examples of most of the categories [23, 26].

In the general theory of surface mapping class groups, pseudo-Anosov automorphisms play the major role. However, every Heegaard splitting mapping class group that has been classified as of today is generated by reducible automorphisms. This make reducible elements particularly important and suggests the following:

4. **Problem.** Is the mapping class group of every (unstabilized) Heegaard splitting generated by reducible automorphisms of the Heegaard surface?

The PI expects that the techniques developed in Objective 1 will contribute to the solution of this problem. In the mean time, the characterization of reducible automorphisms of strongly irreducible Heegaard splittings has considerable room for improvement:

Every unstabilized Heegaard splitting can be decomposed into a collection of incompressible surfaces and strongly irreducible Heegaard surfaces for their complementary pieces [42]. The characterization of reducible automorphisms of these strongly irreducible splittings determine a class of surfaces along which the three-manifold can be further decomposed. Understanding how these two decompositions interact would provide a very thorough hierarchical characterization of reducible Heegaard splitting automorphisms.

5. **Objective.** The PI will determine a hierarchical characterization of all reducible automorphisms of unstabilized Heegaard splittings.

Objective 5, and an answer to Problem 4, would lend valuable insight into the Goeritz problem.

3. Finite covers and immersed surfaces

One of the most exciting recent developments in low dimensional topology has been the proof of the Virtually Fibered Conjecture [1]. This conjecture was originally proposed by Thurston as an approach to proving his conjecture that pieces in the canonical (topological) decompositions of three-dimensional manifolds admit homogeneous geometric structures. The Geometrization Conjecture was proved in 2002 by Grisha Perelman. (The Poincaré Conjecture is a Corollary of Geometrization.) However, the Virtually Fibered Conjecture remained open until early 2012.

8. **Theorem** (Virtual Fibered Conjecture, Agol '12 [1])**.** *Every hyperbolic three-dimensional manifold M is virtually fibered, i.e. M admits a finite covering map $c : N \to M$ from a surface bundle N.*

As a number of topologists have noted, this suggests a program for the classical problem of classifying all three-dimensional manifolds: Every hyperbolic surface bundle is determined by a pseudo-Anosov map, and such maps are fairly well understood via Thurston-Nielsen theory [7]. To classify three-dimensional manifolds, one must determine all the different manifolds covered by each surface bundle, then decide when two different surface bundles cover the same manifold.

9. **Motivating Problem.** Classify three-dimensional manifolds by characterizing finite covers of manifolds by surface bundles.

Again, we do not propose to completely solve this problem, but to make progress that will shed light on its eventual solution. Agol's proof builds off of machinery developed by Wise [46] using $CAT(O)$ cube complexes and a recent Theorem by Kahn-Markovic [18]. Kahn-Markovic show that every hyperbolic three-dimensional manifold admits a large number of immersed incompressible surfaces. Ideally, one would hope that one of these surfaces would be the image of a fiber of a surface bundle N that covers M. Instead, Agol uses this large number of immersed incompressible surfaces to build a $CAT(O)$ cube complex with the same fundamental group as M. Wise's work [46] implies that such a fundamental group has a property called LERF (Locally Extendable Residually Finite) which implies that M is Virtually Fibered.

In addition to this recent line of work, finite covers of three-dimensional manifolds have also received a lot of attention within gometric group theory circles, particularly among those interested in expanders [34]. Much of this was sparked by Lackenby's results on the growth of Heegaard genus under towers of finite covers [31].

All this suggests that the study of immersed surfaces and their relationship to finite covers will be a major future direction for the field of three-dimensional topology.

Proposed Research Plans. The topology of immersed and embedded surfaces should provide a bridge for understanding how finite covers are determined by the large-scale geometry of a given three-dimensional manifold. As described in the previous section, the topology of embedded surfaces is closely connected to Biringer-Souto's characterization of generic three-manifolds [4]. Adapting embedded surface techniques to immersed surfaces should suggest connections between this characterization and finite covers.

The simplest example of this is a manifold constructed by gluing together two handlebodies by a high distance map. As noted above, the low genus Heegaard splittings of such a manifold have been classified using topological means by Scharlemann-Tomova [43] and using geometric means by Hass-Thompson-Thurston [13]. However, the geometric argument allows the surface to be immersed. (This is necessary in order to guarantee a harmonic representative of the surface.) Therefore, it seems very likely that the proof can be adapted to use an immersed surface that is the image of a Heegaard surface under a finite cover, and not necessarily homotopic to a Heegaard surface for the original manifold.

6. **Objective.** The PI will generalize Hass-Thompson-Thurston's proof to immersed surfaces defined by Heegaaard surfaces of finite covers

This should produce a very nice characterization of the low degree covers of such a manifold. In particular, we expect the minimal genus Heegaard surfaces of all low degree covers to be lifts of the high distance Heegaard surface.

While the geometric arguments in [13] are very intuitive, the combinatorial/topological arguments have proved to be more precise and more straightforward to generalize. In particular, the PI's topological interpretation [20] of Hass-Thompson-Thurston's argument led to a number of such generalizations [23, 26]. It would therefore be useful to adapt the immersed surface generalization to the topological setting.

5. **Problem.** Is there a combinatorial explanation of this generalization of Hass-Thompson-Thurston?

For example, a topological proof should make it immediately possible to generalize Objective 6 to gluings of larger numbers of pieces, as in Objective 3. For embedded surfaces, the topological proof uses the fact that an embedded surface intersects a continuous family of Heegaard surfaces in a family of embedded loops that define a path in the curve complex. This is not the case for an immersed surface, since the intersections will also be immersed.

The geometric proof uses the lengths of the intersections to find a lower bound on the area of the immersed or embedded surface, which is turned into a genus bound using the Gauss-Bonnet theorem. The geometric argument is therefore indifferent to whether the surface is immersed or embedded. The strong connection between geometry and topology in dimension three suggests that there should be a similar combinatorial/topological argument, but this will require a careful understanding of immersed loops in surfaces.

7. **Objective.** The PI will develop a theory of immersed curves in surfaces in analogy with the complex of (embedded) curves, in order to characterize immersed surfaces in manifolds with high distance Heegaard splittings.

One of the fundamental ideas that connects the topology of embedded surfaces to the geometry of the ambient manifold is the notion of thin position. This idea, first proposed by Gabai in the context of knots [10], can be thought of as a way of finding Morse functions that efficiently represent a given topological space. Gabai's original idea has been generalized in a number of ways [14, 35, 42], and in particular, thin position is the machinery behind most of the results discussed in the first component of this proposal.

The PI has shown that most generalizations of thin position can be described by an axiomatic framework based on a cell complex called the *complex of surfaces* [21]. In this complex, vertices are embedded surfaces and each edge connects a surface to a new surface that results from compressing it. This can be generalized fairly naturally to immersed surfaces by constructing a complex in which vertices are homotopy classes of immersed surfaces and each edge connects a surface to the immersed surface that results from replacing an immersed annulus in the original surface with two parallel immersed disks.

The complex of immersed surfaces does not have all the structure of the embedded surface complex; for example, compressions of embedded surfaces split into two types, depending on which side of the surface they are on. For immersed surfaces, there is no way to define which "side" the two immersed disks are added on.

In the complex of (embedded) surfaces, sidedness is used to determine when a path in the complex corresponds to a sequence of disjoint surfaces (roughly, the level sets of a possible Morse function.) For the immersed surface complex, there is no such condition. However it is almost immediate from the definition that every path in the immersed surface complex for a manifold M defines a second three-dimensional manifold N and a map $f : N \to M$ such that the immersed surfaces in M are the images under f of a sequence of disjoint surfaces in N.

Standard thin position results translate to combinatorial statements about paths in the complex of surfaces, and such statements can similarly be translated to understand the induced maps f. However, this translation is not immediate.

6. **Problem.** How do (local) properties of paths in the complex of immersed surfaces correspond to properties of the induced map f?

8. **Objective.** The PI will work with Trent Schirmer (a postdoc at OSU) to determine a local condition on paths in the complex of immersed surfaces to calculate the index of the induced map f and to determine when f is a covering map.

4. Applications to data analysis

One of the key problems in the analysis of the large, high dimensional data sets is the classification and description of *clusters* - subsets of the data in which points are more closely related to each other than to points outside the cluster. In bioinformatics, for example, finding clusters of genes based on when they are expressed suggests which genes are involved in the same biological processes. Similarly, in chemometrics, clustering can identify molecules with similar structures based on information such as spectroscopy data.

Clustering algorithms fall into a number of categories determined by the assumptions they make about the distribution of data and their approach to searching for clusters. In general, one expects a topological approach to data analysis to be more robust and impervious to noise than traditional rigid, geometric approaches. Two examples of more topological methods are spectral clustering [8] and Carlsson-Memoli's multi-parameter clustering [6].

One can translate an abstract data set into a graph with weighted edges such that the weight corresponds to how close two points are in a carefully chosen metric. In this setting, clusters are subgraphs that can be separated from the rest of the graph by cutting relatively few edges. (For example, this is the basis of spectral clustering.)

Thin position relates to a similar problem for three-dimensional manifolds, namely searching for topologically efficient decompositions of three-dimensional spaces. In [29], the PI translated the basic ideas from thin position into a data clustering algorithm. The resulting algorithm is gradient-like in the sense that it begins with an ordering of the vertices of the graph, defines a "width" of the ordering, then looks for ways to find "thinner" orderings. Once there are no more possible improvements, there is a simple criteria that decides if the first k vertices should be considered a cluster for any $k < N$. This algorithm, named TILO/PRC (Topologically Intrinsic Lexicograph Ordering/Pinch Ratio Clustering) has been implemented by Doug Heisterkamp (a computer scientist at OSU) and the effectiveness of this approach on both real and synthetic data sets has been demonstrated [15].

In the three-dimensional setting, thin position defines a sequence of surfaces such that the locally minimal surfaces (with respect to genus) are incompressible [42]. In the TILO/PRC algorithm, the surfaces are replaced by the set of edges from the first k vertices to the last $N - k$ vertices and genus is replaced by the sum of the edge weights. Therefore, one would expect local minima with respect to the edge weight sum to be structurally significant. The PI has shown [29] that the clusters defined by local minima satisfy a very strong condition analogous to being the complementary components of an incompressible surface. Such a cluster is called a *pinch cluster*.

This suggests that many of the techniques that have been developed in three-dimensional topology for understanding incompressible surfaces could have strong implications for data analysis. Furthermore, Rubinstein [41] and Bachman [3] have shown that the locally maximal surfaces related to thin position can be thought of as topological/discrete versions of higher index (geometric) minimal surfaces (such as soap films). This suggests that thin position could be used to adapt machinery from the study of minimal surfaces into the TILO/PRC framework.

Proposed Research Plans: Clustering algorithms can be evaluated with a data set in which points of data fall into a small number of classes, by checking how closely the clusters determined by the algorithm correspond to the actual classes. The classic example is the IRIS data set, which consists of measurements taken from 150 different iris plants in three different species. A good clustering algorithm should be able to split the data points into three sets such that each set contains all (or mostly) the same species.

10. **Motivating Problem.** Create a clustering algorithm that significantly outperforms existing algorithms on a wide range of data sets.

One of the key ideas in thin position is that in order to understand the local minima (which in our setting define the clusters) one must control the local maxima. In the topological setting,

each local maximum corresponds to a Heegaard splitting for a complementary component of the local minima. As noted above, distance in the curve complex is a very effective measure for understanding both the topology and the geometry of these components. In addition to the geometric picture suggested by Namazi-Souto [37], Hartshorn [12] has shown that low genus incompressible surfaces cannot cut across regions with high distance Heegaard splittings. (Theorem 5 is a generalzation of this.) This implies that if one finds a thin position with a sufficiently high distance maximum then any other thin position must either contain the same complementary piece or must have much larger local minima.

Because the TILO algorithm progressively improves an initial ordering, there is always the possibility that the final strongly irreducible ordering is a local minimum far from the global minimum. Early tests by the PIs [15] found examples in which the difference in accuracy between different initial orderings of the same data set varied from 1% to 70%. Because it is impractical to test every possible ordering of a data set with thousands of points, it would be very useful to have a way of measuring how stable any given ordering is and certify that a given strongly irreducible ordering is close to the absolute minimum. A definition in the data setting equivalant to curve complex distance, and a result along the lines of Hartshorn's would allow one to determine when a given ordering is reasonably efficient

9. **Objective.** The PI will adapt the notion of curve complex distance to a measure of the stability of clusters.

The topological proof of Hartshorn's Theorem [12] requires a careful study of the cross sections in the region defined by a local maximum. The PI proposes to expand the similarity graph to a simplicial complex in order to determine a more precise definition of cross sections, and compare the cross sections defined by different orderings of the vertices. Because this problem has been thoroughly studied in the topology setting, this should be a reasonably straightforward translation of ideas.

The TILO algorithm reduces a given ordering by applying a sequence of permutations called *shifts* and at each step it makes a number of choices. Depending on the order in which the shifts are made, the final strongly irreducible orderings can be vastly different. Some steps can be done in different orders to produce the same result, but there are also fork points in which the decision of which shift to perform is irrevocable. In the topological setting, the fork points where the final ordering is determined are characterized by what are called *index-two* surfaces. These have been studied by Bachman [3] and by the PI [21], and play a large role in understanding common stabilizations and isotopies between different representatives of the same Heegaard surface.

In the data setting, an index-two surface would correspond to an ordering in which there are two shifts that reduce the width, but carrying out either one of the shifts would interfere with the other one. The current implementation of the TILO algorithm finds a single sequence of shifts starting from an initial ordering and chooses the shifts greedily - it always uses the shift that immediately reduces the width as much as possible. On the other hand, understanding the fork points could suggest either an algorithm to efficiently carry out multiple sequences of shifts, or a better (non-greedy) heuristic for choosing a single sequence of shifts.

10. **Objective.** The PI will adapt the notion of index-two thin position to determine a heuristic for an optimal thinning strategy in TILO.

The goal will be to adapt the complex of surfaces mentioned above to develop an abstract framework for studying the entire set of orderings of vertices and how they are related by shifts. Because this theory has been thoroughly developed in the topology setting, it should be relatively straightforward to translate into the data setting. The difficulty will be in interpreting the results and applying them to determine heuristics for the TILO algorithm.

The PI and Doug Heisterkamp are planning to implement a version of TILO/PRC for streaming data by considering a similarity graph that changes with time. In particular, each vertex will be given a Gaussian weight function with respect to time and the edge weights will be scaled by the weights at their endpoints. The difficulty will be to effectively interpret the output of the TILO/PRC algorithm as it changes with time.

Because index-two surfaces in the topological setting determine how different Heegaard surfaces are related to each other, the equivalent notion in the data setting should lend insight into how different clustering partitions are related. In the analysis of streaming data, the ordering or branched ordering changes automatically over time and these changes should be determined by phenomenon closely related to index-two surfaces.

11. **Objective.** The PI will adapt the notion of index-two thin position to interpret the results of the time series version of TILO/PRC.

The TILO/PRC algorithm allows the structure of the data to completely determine the size (and possibly also the number) of clusters. However, it is often desirable to partition the data into equally sized or fixed sized subsets. For example, to find a small number of representatives of the entire set, one might decompose the set into clusters of a fixed size n, then choose one representative from each cluster.

The thin position analogy suggests that we can think of this from the perspective of minimal surfaces/soap films. Because the clusters in TILO/PRC are not a fixed size, the separations between them behave like a film bounded by a wire loop, in which air can pass from one side to the other and the volume is not fixed. Such a film is, in some sense, convex on each side, and the definition of a pinch cluster [29] is a combinatorial version of this. To find fixed sized clusters, we should think of soap bubbles, in which the amount of air inside the bubble is fixed. The resulting surface is convex in one direction, but generally concave in the other. The definition of pinch cluster suggests an analogous notion for subsets of a data set.

12. **Objective.** The PI will develop an algorithm for balanced clustering, in which cluster sizes are fixed or constrained.

Such an algorithm will be similar to the the TILO/PRC algorithm, but rather than choosing a linear ordering on the data points, this algorithm will cut the data into some fixed number of subsets and will only consider the boundary of each set (i.e. the sum of weights of edges from a vertex in the set to a vertex outside). Sets will be allowed to trade vertices if the trade reduces the boundary of both sets.

For this objective, it is not immediately clear what model from the topological setting is most relevant. In order to develop heuristics in the data setting, the PI will translate this problem into one or more topological problems and look for solutions based on new or existing theory. These theoretical solutions will then be used as models for understanding the problem in the data setting.

Another approach to data analysis that has recently come out of topology is the notion of Persistent Homology [9]. In the topological setting, homology and thin position have played complementary roles. In particular, thin position has proved very useful for discovering topological structure that is not seen by homology. (For example, Heegaard surfaces are all homologically trivial but can be distinguished using thin position.) In three-dimensional topology, thin position and homology theory together provide a very thorough picture of the structure of a three-dimensional space. It is likely that the same will prove to be true in the data setting.

13. **Objective.** The PI will compare the TILO/PRC approach to other topological approaches to data analysis such as Persistent homology.

In particular, circular thin position [35] is a recently developed generalization of thin position that takes into account non-trivial homology in a three-dimensional manifold. Roughly speaking, this would correspond in the data setting to choosing a circular rather than linear ordering of the data points that wraps around a non-trivial cohomology cycle. In general, it should be possible to adapt ideas from circular thin position in order to replace the tree structure of K-branched PRC with a more general graph determined by the persistent homology of a data set. This has the potential to increase the accuracy of the final cluster structure determined by such an algorithm.

The PI will also investigate applying the TILO/PRC algorithms to persistence complexes, as opposed to just weighted graphs. A persistence complex can be thought of as a graph (or a simplicial complex) in which the edges, faces and higher dimensional cells are introduced one at a time, starting with the shortest edges. The graph slowly grows from a set of isolated vertices to the final final graph, and TILO can be applied at each stage, starting from an ordering defined by the previous stage. This is reminiscent of the streaming data problem discussed in Objective 11, and would result in a final ordering that is less dependent on the initial (essentially random) ordering of the vertices and more dependent on the data structure.

The PI and Doug Heisterkamp are currently working with Barry Lavine (a chemist at OSU) to integrate TILO/PRC into a genetic feature selection algorithm developed previously by Lavine. Lavine has used the original algorithm to determine which wavelengths of light are best suited to distinguishing certain classes of chemicals [32]. The genetic algorithm selects different sets of features (wavelengths) and tests how well these distinguish the different classes of chemicals. This measure is called a *fitness function* and the different feature sets are combined and modified to form new sets, with preference given to feature sets with higher fitness values.

The PI has developed a fitness function based on TILO/PRC, using the classification accuracy and a measure called the pinch ratio for the final measure. Preliminary tests suggest that this approach should be much more effective than Lavine's original fitness function. We plan to implement this new fitness function and use it to re-analyze data from Lavine's past experiments.

The PI plans to initiate similar collaborations with scientists in other departments at OSU. The University has recently created an interdisciplinary bioinformatics program and the PI is involved in the internal competition for an IGERT proposal headed by Barry Lavine to begin a chemometrics program. Because of these programs, a number of students and faculty across the University have become interested in understanding both sides of the science/informatics connection.

14. **Objective.** The PI will initiate collaborations with the BioInformatics and Chemometrics programs being developed with OSU as well as with scientists at OSU and elsewhere.

5. Broader Impacts

The PI is currently mentoring a third-year graduate student (Pengcheng Xu), is a member of three additional dissertation committees and has organized a number of conferences with large participation by graduate students. Two of these conferences, "Triangulations, Invariants and Geometric Structures" (2010) and "The Redbud Topogy Conference" (2012) were held at Oklahoma State University, providing important opportunities for geographically isolated graduate students. He will continue to recruit students who are interested in pure and applied topology, and to organize conferences and seminars aimed at a broad audience.

The PI maintains a research blog *Low dimensional topology* [19], which discusses recent research in pure and applied topology. This blog averages over 200 views per day and a recent series on the applications of topology to data analysis has attracted readers from both pure mathematics and applied data mining.

The applied component of this project will create connections between mathematics/computer science and the physical sciences, develop the infrastructure for data rich science research and create new opportunities within pure mathematics and computer science.

The applications of topological methods to data analysis has potentially huge ramifications for the advancement of data-rich science such as bioinformatics and chemometrics, and will help demonstrate the benefits of pure math research to a broad audience. Moreover, by creating direct connections between pure mathematics and applied science, the project will increase the number of employment opportunities for students on the pure research side, which will help attract students into mathematics and computer science.

It should be noted that OSU has a substantial population of Native American and other underserved minorities, and Oklahoma is geographically isolated from the academic centers of the country. Through involvement in the PI's research, mathematically talented students at OSU will have the opportunity to develop their talents, increase their visibility and confidence and prepare themselves for further success in mathematics and science.

6. Results of prior support

The PI is currently supported by NSF grant DMS-1006369, "The geometry and topology of Heegaard splittings". The project supported by this grant has developed the foundation for all three components of the current proposal.

This grant will end August, 2013. During the first two years of the grant DMS-1006369, The PI has produced eleven papers.

The first four papers developed the Pi's axiomatic thin position machinery in order to prove a number important results about Heegaard splittings. The first of these proves a bound on the size of Heegaard trees, namely that any two Heegaard splittings of the same three-manifold, with genera $p \geq q$ have a common stabilization of genus at most $\frac{3}{2}p + 2q - 1$. This greatly improved an existing upper bound by Rubinstein-Scharlemann [39].

(1) *An upper bound on common stabilizations of Heegaard splittings*, preprint, arXiv:1107.2127.

(2) *Mapping class groups of Heegaard splittings of surface bundles*, preprint, arXiv:1201:2628.

(3) *One-sided and two-sided Heegaard splittings*, preprint, arXiv:1112.0471.

(4) *Mapping class groups of once-stabilized Heegaard splittings*, preprint, arXiv:1108.5302.

The next five papers explore connections between the theory of Heegaard splittings and other approaches to three-manifold topology. These include using the theory of geometric limits to study automorphisms of handlebodies and using mapping class groups of Heegaard splittings to characterize open book decompositions.

(1) *Extending pseudo-Anosov maps to compression bodies* (with I. Biringer and Y. Minsky), to appear in *Journal of Topology*, arXiv:1011.0021.

(2) *The space of Heegaard splittings* (with D. McCullough), to appear in *Crelle*, arXiv:1011.0702.

(3) *Layered models for 3-manifolds*, in *Topology and Geometry in Dimension Three*, Contemporary Mathematics, 565 (2011), 43–54, arXiv:1011.6343.

(4) *The coarse geometry of the Kakimizu complex* (with Roberto Pelayo and Robin Wilson), preprint, arXiv:1204.0530.

(5) *Heegaard splittings and open books*, preprint, arXiv:1110.2142.

The final two introduced the application of thin position to the analysis of large data sets. The first of these papers develops the theoretical framework, while the second demonstrates its effectiveness compared to existing data analysis algorithms.

(1) *Topological graph clustering with thin position*, preprint, arXiv:1206.0771.

(2) *Pinch Ratio Clustering from a Topologically Intrinsic Lexicographic Ordering*, Doug Heisterkamp and Jesse Johnson, submitted to *Proceedings of the IEEE International Conference on Data Mining (ICDM 2012)*, December 10-13, 2012.

The PI has also been a CO-PI on two conference grants within the past five years: DMS-0602638, "Geometric Topology in Three and Four Dimensions" supported a three day conference at UC Davis in Summer 2009 and DMS-1148725, "Redbud Topology Conference" supported a weekend conference at Oklahoma State University in Spring, 2012. Both grants were used primarily to fund graduate student travel, allowing a large number of graduate students to attend each conference.

References

1. Ian Agol, *The virtual haken conjecture*, preprint (2012), arXiv:1204.2810. Appendix by Daniel Groves and Jason Manning.
2. D. Bachman, *Stabilizations of Heegaard splittings of sufficiently complicated 3-manifolds (Preliminary Report)*, preprint (2008), arXiv:0806.4689.
3. David Bachman, *Topological index theory for surfaces in 3-manifolds*, Geom. Topol. **14** (2010), no. 1, 585–609. MR 2602846 (2011f:57042)
4. Ian Biringer and Juan Souto, *A finiteness theorem for hyperbolic 3-manifolds*, preprint (2009), arXiv:0901.0300.
5. Ryan Blair, Marion Campisi, Jesse Johnson, Scott A. Taylor, and Maggy Tomova, *Bridge distance, heegaard genus and exceptional surgeries*, preprint (2012).
6. Gunnar Carlsson and Facundo Mémoli, *Multiparameter hierarchical clustering methods*, Classification as a tool for research, Stud. Classification Data Anal. Knowledge Organ., Springer, Berlin, 2010, pp. 63–70. MR 2722123
7. Andrew J. Casson and Steven A. Bleiler, *Automorphisms of surfaces after Nielsen and Thurston*, London Mathematical Society Student Texts, vol. 9, Cambridge University Press, Cambridge, 1988. MR 964685 (89k:57025)
8. W. E. Donath and A. J. Hoffman, *Lower bounds for the partitioning of graphs*, IBM J. Res. Develop. **17** (1973), 420–425. MR 0329965 (48 #8304)
9. Herbert Edelsbrunner, David Letscher, and Afra Zomorodian, *Topological persistence and simplification*, Discrete Comput. Geom. **28** (2002), no. 4, 511–533, Discrete and computational geometry and graph drawing (Columbia, SC, 2001). MR 1949898 (2003m:52019)
10. David Gabai, *Foliations and the topology of* 3*-manifolds. III*, J. Differential Geom. **26** (1987), no. 3, 479–536. MR 910018 (89a:57014b)
11. L. Goeritz, *Die Abbildungen der Brezelfläche und der Volbrezel vom Gesschlet 2*, Abh. Math. Sem. Univ. Hamburg **9** (1933), 244–259.
12. Kevin Hartshorn, *Heegaard splittings of Haken manifolds have bounded distance*, Pacific J. Math. **204** (2002), no. 1, 61–75. MR 1905192 (2003a:57037)
13. Joel Hass, Abigail Thompson, and William Thurston, *Stabilization of Heegaard splittings*, Geom. Topol. **13** (2009), no. 4, 2029–2050. MR 2507114 (2010k:57044)
14. Chuichiro Hayashi and Koya Shimokawa, *Thin position of a pair (3-manifold, 1-submanifold)*, Pacific J. Math. **197** (2001), no. 2, 301–324. MR 1815259 (2002b:57020)
15. Doug Heisterkamp and Jesse Johnson, *Pinch Ratio Clustering from a Topologically Intrinsic Lexicographic Ordering*, preprint (2012).
16. John Hempel, *3-manifolds as viewed from the curve complex*, Topology **40** (2001), no. 3, 631–657. MR 1838999 (2002f:57044)
17. William Jaco, Hyam Rubinstein, and Stephan Tillmann, *Minimal triangulations for an infinite family of lens spaces*, J. Topol. **2** (2009), no. 1, 157–180. MR 2499441 (2010b:57016)
18. Vladimir Markovic Jeremy Kahn, *Immersing almost geodesic surfaces in a closed hyperbolic three manifold*, preprint (2009), arXiv:0910.5501.
19. Jesse Johnson, *Low dimensional topology*, http://ldtopology.wordpress.com/.
20. ———, *Bounding the stable genera of Heegaard splittings from below*, J. Topol. **3** (2010), no. 3, 668–690, arXiv:0807.2866. MR 2684516
21. ———, *Computing isotopy classes of Heegaard splittings*, preprint (2010), arXiv:1004.4669.
22. ———, *An upper bound on common stabilizations of Heegaard splittings*, preprint (2011), arXiv:1107.2127.
23. ———, *Heegaard splittings and open books*, preprint (2011), arXiv:1110.2142.
24. ———, *Layered models for closed 3-manifolds*, Topology and geometry in dimension three, Contemp. Math., vol. 560, Amer. Math. Soc., Providence, RI, 2011, pp. 43–54. MR 2866922
25. ———, *Mapping class groups of once-stabilized Heegaard splittings*, preprint (2011), arXiv:1108.5302.
26. ———, *One-sided and two-sided Heegaard splittings*, preprint (2011), arXiv:1112.0471.
27. ———, *Handlebody Filling and the Heegaard Tree*, Joint Mathematics Meetings 2012, special session on Hyperbolicity, 2012.
28. ———, *Mapping class groups of Heegaard splittings of surface bundles*, preprint (2012), arXiv:1201.2628.
29. ———, *Topological graph clustering with thin position*, preprint (2012), arXiv:1206:0771.
30. Jesse Johnson and Hyam Rubinstein, *Mapping class groups of Heegaard splittings*, preprint (2006), arXiv:math.GT/0701119.

31. Marc Lackenby, *Heegaard splittings, the virtually Haken conjecture and property* (τ), Invent. Math. **164** (2006), no. 2, 317–359. MR 2218779 (2007c:57030)
32. B.K. Lavine, C.E. Davidson, and A.J. Moores, *Genetic algorithms for spectral pattern recognition*, Vibrational Spectroscopy **28** (2002), no. 1, 83–95.
33. Tao Li, *Heegaard surfaces and the distance of amalgamation*, Geom. Topol. **14** (2010), no. 4, 1871–1919. MR 2680206 (2011j:57027)
34. D. D. Long, A. Lubotzky, and A. W. Reid, *Heegaard genus and property τ for hyperbolic 3-manifolds*, J. Topol. **1** (2008), no. 1, 152–158. MR 2365655 (2008j:57036)
35. Fabiola Manjarrez-Gutiérrez, *Circular thin position for knots in S^3*, Algebr. Geom. Topol. **9** (2009), no. 1, 429–454. MR 2482085 (2010i:57020)
36. Yair Minsky, *The classification of Kleinian surface groups. I. Models and bounds*, Ann. of Math. (2) **171** (2010), no. 1, 1–107. MR 2630036 (2011d:30110)
37. Hossein Namazi and Juan Souto, *Heegaard splittings and pseudo-Anosov maps*, Geom. Funct. Anal. **19** (2009), no. 4, 1195–1228. MR 2570321 (2011a:57035)
38. K. Reidemeister, *Zur dreidimensionalen Topologie*, Abh. Math. Sem. Univ. Hamburg **11** (1933), 189–194.
39. Hyam Rubinstein and Martin Scharlemann, *Comparing Heegaard splittings of non-Haken 3-manifolds*, Topology **35** (1996), no. 4, 1005–1026. MR 1404921 (97j:57021)
40. J. H. Rubinstein, *Polyhedral minimal surfaces, Heegaard splittings and decision problems for 3-dimensional manifolds*, Geometric topology (Athens, GA, 1993), AMS/IP Stud. Adv. Math., vol. 2, Amer. Math. Soc., Providence, RI, 1997, pp. 1–20. MR 1470718 (98f:57030)
41. J. Hyam Rubinstein, *Minimal surfaces in geometric 3-manifolds*, Global theory of minimal surfaces, Clay Math. Proc., vol. 2, Amer. Math. Soc., Providence, RI, 2005, pp. 725–746. MR 2167286 (2006g:57038)
42. Martin Scharlemann and Abigail Thompson, *Thin position for 3-manifolds*, Geometric topology (Haifa, 1992), Contemp. Math., vol. 164, Amer. Math. Soc., Providence, RI, 1994, pp. 231–238. MR 1282766 (95e:57032)
43. Martin Scharlemann and Maggy Tomova, *Alternate Heegaard genus bounds distance*, Geom. Topol. **10** (2006), 593–617 (electronic). MR 2224466 (2007b:57040)
44. James Singer, *Three-dimensional manifolds and their Heegaard diagrams*, Trans. Amer. Math. Soc. **35** (1933), no. 1, 88–111. MR 1501673
45. Michelle Stocking, *Almost normal surfaces in 3-manifolds*, Trans. Amer. Math. Soc. **352** (2000), no. 1, 171–207. MR 1491877 (2000c:57045)
46. Daniel Wise, *The structure of groups with a quasiconvex hierarchy*, preprint (2011), http://www.math.mcgill.ca/wise/papers.html.