Brandeis University

Undergraduate Thesis in Computer Science

Graph Isomorphism, Reconstruction and the Cycles Invariant

Grady Ward

supervised by
Prf. James Storer

February 29, 2016

Contents

Introduction

We can't choose what interests us, but we do get to choose what we pursue. When I asked Professor Storer to supervise me on a thesis on the graph isomorphism problem, he was hesitant. He only acquiesced when I persuaded him that my future was secure with a fantastic job, and that my primary objective was the pursuit (and possible rediscovery) of questions that were of nothing but personal interest to me. In many respects, his skepticism proved well founded. This work has been incredibly challenging, both in that the body of existing work on GI is so large, and in that finding niches of it that are promising and not yet fully explored is difficult. Over the past year I have poured my time and energy into this project, and have found it unbelievably energizing to do so. I have been thrilled to find interesting properties in problems surrounding GI, and have had an equal number of frustrations in finding that my results had been previously discovered. I would like to thank Professor Storer for the initial bout of skepticism about this project, as it shaped this project and experience for the better. It has kept me on track to focus on my real goal for the semester, which was to grow. I have learned advanced techniques in GPU calculation, proof techniques in abstract algebra, and research and documentation techniques. My skill set has been broadened by a project which has deeply challenged me and always kept me on my toes.

Chapter 1

Definitions and Syntax

1.1 Graphs

This report describes undirected, unlabeled graphs with no self-loops or multi-edges. Such graphs are represented by an *adjacency matrix*: a square, symmetric, binary matrix with zeros along the diagonal. We will use G to refer to a graph, and A to refer to an adjacency matrix of the graph. When we use A, it will not refer to a specific adjacency matrix, as most graphs can be represented by many matrices.

We will refer to the set of vertices of a graph as $V(G)$, and the set of edges of a graph as $E(V)$. Graphs have N vertices and M edges, and it is frequently used without clarification that $M \in [0, \frac{1}{2}(N)(N-1)]$. When discussing specific edges, tuples are symmetric: $(v_1, v_2) = (v_2, v_1)$.

A graph's *complement* is a new graph over the same set of vertices, but where adjacency and non-adjacency are inverted. In formal terms, the graph H is G's inverse if $V(H) = V(G)$ and $(v_1, v_2) \in E(H) \leftrightarrow (v_1, v_2) \notin E(G)$.

1.1.1 Representations, Labeling, Matrices

An important distinction in this report is on the *representation* of graphs. Most graphs can be represented by distinct adjacency matrices, but changes to representation do not change the fundamental structure of the graph that the matrices represent. Different representations of graphs are akin to different labelings of the graph: neither mutate structure, and neither should factor into our algorithms. Throughout this report that when we discuss a set of graphs, or an algorithm over graphs, we will treat graphs as objects which denote structure, and will in every way be blind to their representation. Thus when we refer to a graph, we are referring to all of its representations. When we intend to discuss a graph within some physical reality of its representation (for example, when determining whether two given adjacency matrices refer to the same graph) we will use the term

graph instance. This is not a semantic choice, it has important implications (particularly around ideas of random graph models).

1.2 Graph Isomorphism and Automorphism

Two graph instances G and H are *isomorphic* if there exists a mapping M between $V(G)$ and $V(H)$ such that

$$\forall_{a,b \in V(G)}(a,b) \in E(G) \leftrightarrow (M(a), M(b)) \in E(H)$$

. If an isomorphism exists between two graph instances, then the two instances represent the same graph; they have the same structure. An isomorphism preserves all adjacencies and all non-adjacencies, and the existence of an isomorphism between instances proves that they are the same graph. It may be possible for multiple isomorphisms to exist between two graph instances, but we are generally only concerned with the existence of such a mapping. We will use the notation $Iso(G,H)$ to be shorthand for a boolean predicate describing the existence of such a mapping.

The question of whether or not graph isomorphism as a descision problem (GI) can be computed in polynomial time is an open question in theoretical computer science. The problem is known to be solveable in quasi-polynomial time, though no convincing arguments have placed it in NP-complete nor in P.

An *automorphism* is a mapping of the set of vertices of a graph onto itself ($V(G)$ to $V(G)$) which preserves adjacency and non-adjacency. If an automorphism M maps every element of $V(G)$ to itself, the automorphism is called the *trivial automorphism*. Though it will be taken as granted, the set of all valid automorphisms for a graph G forms a group. This group has at least one element (the identity element is the identity isomorphism), but may have as many as $N!$ elements. This group will be referred to as $Aut(G)$, and the operation over the group is understood to be the *followed by* operation.

1.2.1 Graph Invariants

A *graph invariant* is an ordered property of a graph which remains the same irrespective of the representation or labeling of the graph. More specifically, an algorithm or property is a graph invariant only if it produces output which is stable across all instances of the same graph. A graph invariant $I(G)$ can allow us to conclude that two graphs (G_1, G_2) are *not* the same graph if $I(G_1) \neq I(G_2)$. However, it is distinctly limited, in that the converse does not hold (i.e. it is possible for non-isomorphic graph instances to share a value for a graph invariant).

1.2.2 Discriminatory Power

A graph invariant is *discriminating* if it can, with a certain probability, distinguish two non-isomorphic graphs as non-isomorphic. For example, an example of a graph invariant that is not very discriminatory is the vertex count of a graph. Two graphs are certainly not isomorphic if they differ in their vertex count, however, many graphs which are not isomorphic have the same vertex count. In contrast, the chromatic polynomial of a graph is a highly discriminatory graph invariant, as the odds of having two non-isomorphic graph instances agree on their chromatic polynomial is relatively low.

To formalize this notion, we will discuss discriminatory power with a specific probabilistic meaning. A graph invariant I discriminates at a level α for N vertices and M edges if selecting two graphs G and H at random from the set of all graphs with N vertices and M edges:

$$P(I(G) = I(H) \wedge \neg Iso(G, H)) \leq \alpha$$

What we will find is that we can frequently discuss alpha as a function of M and N. Later in this report we will discuss how α fits in to a natural definition of a false positive an uncertain test without a false negative rate ($\beta = 0$).

1.2.3 Vertex Similarity

Two vertices are *similar* if there exists a mapping in the automorphism group $Aut(G)$ such that the mapping maps one vertex to the other. Similarity is a transitive property. The vertex set $V(G)$ can be divided up into between 1 and N similar vertex sets, such that all of the vertices in each set are similar, and no pair of sets contains similar vertices. A discussion of these *similar vertex sets* (or SVSs) will be the primary focus of chapter four.

1.2.4 Vertex Invariants

Vertex invariants are numerical properties that we can calculate over a specific vertex within a graph which attempts to identify potentially similar vertex pairs. Similar vertices within a graph will agree on all vertex invariants. However, like graph invariants, vertex invariants can only eliminate the possibility for vertex similarity, they are not sufficient to prove similarity.

Vertex invariants make the computation of graph isomorphism between two graphs markedly easier. Whereas a graph invariant can tell us about whether or not graph instances as a whole might be alike, it does nothing to suggest a proposed mapping between the vertices of the two graph instances. In contrast, a vertex invariant identifies potentially similar vertices not only within a graph, but also between two graphs. A *perfect* vertex invariant

is one for which agreement on the value of the invariant is equivalent to establishing the existence of an automorphism that maps one vertex to the other.

A question discussed later in this report will be about the theoretical implications of a hypothetical perfectly discriminatory vertex invariant, and a couple of ideas that might suggest idealized invariants.

1.3 Cycles Invariant

The focus of this report is an invariant which can function as an invariant over graphs or their vertices. It is called the 'Cycles' invariant, but is sometimes referred to as the 'Paths' Invariant in cases where cycle has other connotations. In either case, we will consistently capitalize to distinguish the invariant from the other denotations.

Cycles as a vertex invariant describes a vector of length P, where the pth entry is the number of closed cycles of length p which pass through the vertex being described. Cycles are allowed to repeat vertices and edges, and can pass back through their place of origin. We are counting cycles which are directional, so $ABCA$ and $ACBA$ are distinct cycles.

If a vertex is the ith row/column of an adjacency matrix A, then the cycles invariant for the vertex is the successive values of

$$Paths(G, v_i, p) = A^p(i, i)$$

for each of the values of p, forming a vector of length P.

The vector generated by this computation is a way of describing the local graph around the vertex v_i. We can consider many ways in which paths reflects a 'reverberation' about the local neighborhood of a vertex, and provides a noisy invariant, which is useful for distinguishing purposes. It will be discussed why later, but we will prove that P will always be strictly less than N, and that further values of computation are not useful.

Extending this vertex invariant to a graph invariant requires no imagination. We simply calculate the paths vertex invariant for every one of the vertices of the graph, and are given back N vectors of length P. We then sort the resultant vectors lexicographically, and arrange them in a NxP matrix. This matrix is a comparable object which is invariant to changes in labeling or representation of G.

1.3.1 Invertibility

Note that if two graphs agree on the paths invariant, their complements (or inverses) also agree on the paths function. This is easily observed through the knowledge that two graphs are isomorphic if and only if their complement graphs are isomorphic.

WAIT COPATHS GRAPHS THIS COULD BE FUCKING HUGE TODO !!! PATHS + INVERSE(PATHS) != PATHS(K) IS THIS FUCKING TRUE?!?!

1.3.2 Running Time

The running time of the paths invariant is the same whether we treat it as a vertex invariant or as a graph invariant. The one difference is whether or not we sort the resulting vectors. It is imperative to calculate A^P even if we are only interested in calculating it for a single vertex. Matrix multiplication can generally be accomplished in sub-cubic time, but for the sake of simplicity and comprehensibility within this paper we will assuem that it is an algorithm that runs in $O(N^3)$ time. This means that the computation of any paths invariant can be accomplished in $O(n^4)$ time. Though this sounds like a lot of time, generally, this is a quick computation. This is made asymptotically better by the fact that efficient algorithms for matrix multiplication are well studied and optimized for a variety of contexts. In a modern context, matrix multiplication can be done efficiently by GPU arrays, and there are quick ways to do multiplication on sparse matrices.

1.3.3 Dealing with Large Numbers

One element of the Cycles invariant that we will consistently need to be cognizant of is the fact that our numbers will grow very quickly, particularly for large values of N. In the worst case, the largest value in the Paths invariant will occur when we have a fully connected graph of size N. The largest value will be TODO as can be quickly shown via TODO.

We can avoid most of the problems associated with this by simply performing regular modulo operations by a large value of 2^K. This preserves thee properties of addition and composition that we are looking for, but does us th favor of maintaining reasonable sized bit arrays. If we break every number into two parts, one divisible by a large power, we can see that any means TODO.

1.3.4 Asymptotic Bit Growth

Though the above simplification can certainly allow us to do our computations in a way that is likely to avoid collisions, if we are discussing properties of the invariant as a whole, it is not acceptable to ignore the possibility of collisions. Thus, when we discuss the cycles invariant within the context of theory, we need to prove that computation of it can be done in linear space. Otherwise, the polynomial aspect of this computation could be misleading away from a gross inefficency in space that masks the the true tosts and limiting factors of the computation.

Proving that the number of bits that the paths function takes is actually not too difficult. TODO Waiting on invertability

1.4 Reconstructability, Determined, Representation

Many of the graph theoretic discussions of the cycles invariant will describe it with respect to other invariants. We have some language to assist these comparisons. We will say that a property of a graph is 'reconstructable' from Cycles if we can calculate the property if we are given the Cycles invariant for the graph. Similarly, a property is determined by Cycles if a set value of cycles allows only zero or one value for the property in question. Reconstructability and determinability differ only in that reconstruct-ability describes a procedure for the conversion, while determinability only claims a valid mapping, and doesn't suggest a mechansim for its computation.

Finally, we will talk about the number of distinct matrices that describe isomorphic graphs as being the 'number of representations' of the graph, or $Reps(G)$.

TODO: What if we have Paths and Cycles (Paths being not closed?) Or like sorted rows/columns of AI???

Chapter 2

Cycles as a Graph Invariant

Cycles is a powerful graph invariant. In this chapter we will discuss properties of cycles that are reconstructable from cycles. We will then show how Cycles is reconstructible from some other graph invariants, which leads us to the conclusion that Cycles is necessarily an incomplete invariant. We will then discuss manual calculations that prove the example of Co-Cycles graphs, and the establishment of small datasets of co-cycles graphs as matrix by which to measure graph invariants. Finally, we will examine the performance of the cycles invariant on different classes of graphs which typically show resistance to classification and differentiation via graph invariants.

2.1 Basic Cycles-Reconstructable Properties

2.1.1 Vertices, Edges, Degree Sequence

From the cycles graph invariant, we can easily deduce the number of vertices (the size of the resulting matrix's first dimension), and the number of edges (the sum of the second column of this matrix, divided by two). We can also deduce the degree sequence, as simply observing the second column of each vertex invariant vector, and sorting the result.

2.1.2 Triangles, Higher Order Polygons

Within the context of a graph, a polygon differs from a cycle in that a cycle is allowed to repeat both edges and vertices, while polygons do not allow repetitions of vertices or edges.

The number of triangles is also easily computed. Since triangles necessarily contain three distinct vertices (in a graph with no self-loops), we know that any cycle of length three on a graph will be a triangle. Thus, the number of triangles which pass through each vertex is simply the third column of the paths invariant matrix, and summing them and dividing by three yields the total number of triangles in the graph.

Things are not so simple for larger polygons. If we think very critically, we can deduce the number of of valid quadrilaterals, by considering the degree sequence of each of the nodes adjacenct to a specific node. We can formalize this notion as follows. TODO Note that this logic requires us to take in an additional piece of information to augment the data that we get from the cycles invariant. Similarly, figuring out higher order polygons can be done, but it requires an awareness of the adjacent values of cycles. More generally, the larger the 'neighbors-paths' supplemental information we require, the further we stray toward giving away the information that fully determines the graph.

However, this is a powerful observation, and fits into a conception of a 'neighbors aware mechanism'. A formal discussion of such mechanism sis discussed in TODO.

2.1.3 Chromatic Polynomial

2.2 Other Forms of Reconstructability

2.2.1 EA Reconstructability

2.2.2 Deck Reconstructability

2.3 Placing Cycles within a Time/Power Tradeoff

2.4 Discrimination on Tough Graph Classes

2.4.1 Background

2.4.2 1-Sparse Graphs

2.4.3 2-Dense Graphs

2.4.4 Miyizaki Graphs

2.5 Imperfection, Co-Cycles Graphs

2.5.1 Discovering Co-Cycles Graphs

2.5.2 Constructing Co-Cycles Graphs

2.5.3 As a Proposed Dataset for Invariant Analysis

2.6 Discriminatory Agreement By N and M

2.6.1 Expectations Borne out of Graph Counts

2.6.2 An Unexpected Dip

Chapter 3

Cycles as a Vertex Invariant

Chapter 4

Cycles and the Reconstruction Conjecture

Chapter 5

Canonical Labeling Using Cycles

5.1 Background

5.2 A Consistent Algorithm

5.3 Time Growth Comparisons to Faster Algorithms

Chapter 6

Random Graph Generators and Automorphisms

Chapter 7

Reflections

Bibliography

[1] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[2] Albert Einstein. *Zur Elektrodynamik bewegter Körper*. (German) [*On the electrodynamics of moving bodies*]. Annalen der Physik, 322(10):891?921, 1905.

[3] Knuth: Computers and Typesetting,
`http://www-cs-faculty.stanford.edu/~uno/abcde.html`