# Subgraphs as a measure of similarity

Josef Lauri

University of Malta josef.lauri@um.edu.mt

## 1 Introduction

Defining the similarity, or distance, between mathematical objects in some class is generally always an important undertaking, and this is no exception for graphs. Ideally we would like to define the similarity between two graphs $G$, $H$ as a parameter which is easy to compute, achieves some maximum value if and only if $G$ and $H$ are isomorphic, and in some sense captures how different $G$ and $H$ are when they are not isomorphic. In a sense, all graphical parameters can be considered candidates for such a similarity measure, but no measure which satisfies all these conditions is known. An easily computable parameter which determines when two graphs are isomorphic would solve the Graph Isomorphism (GI) Problem, one of graph theory's diseases [24]. Easily computable parameters such as the degree sequence and the spectrum do not always distinguish between non-isomorphic graphs. But devising measures which are efficiently computable although not always able to distinguish between non-isomorphic graphs is still an important realm for investigation, especially in applications. A recent example of work in this field (sometimes called *inexact graph matching* [8]) is [9], where the authors derive a hierarchy of similarity measures related to the degree sequence parameter and which can be computed efficiently. In this paper the authors give experimental results obtained by applying their similarity measures to more than four hundred directed graphs representing web-based hypertext structures.

In this chapter we shall focus on measuring the similarity of two graphs in terms of their subgraphs. Complexity considerations and practical use will only be discussed briefly in the last section. The first paper to study this way of measuring similarity or distance between graphs was probably [25]. In this paper, motivated by a question of Vizing, Zelinka defines the the distance $\delta(G, H)$ between two graphs on $n$ vertices as the minimum $k$ such that $G$ and $H$ are both induced subgraphs of a graph on $n + k$ vertices and he shows that $\delta$ is a metric on the set of graphs with $n$ vertices. He also proves the simple result that $G$ and $H$ are induced subgraphs of a graph on at most $n + k$

vertices if and only if they have a common induced subgraph on at least $n-k$ vertices. We shall consider a similarity measure which takes into consideration all induced subgraphs and which is also related to another well-known graph theory disease.

So, how similar can two graphs be? Two graphs are, of course, as similar as they can be when they are isomorphic. However, how much internal structure can two non-isomorphic graphs share? We show what the answer can look like if the measure of common internal structure between the two graphs is taken to be the number of isomorphic subgraphs which they share. We see how this notion is related to the internal symmetries of a graph and that therefore, for most graphs, their internal structure forces them to be very dissimilar to other graphs. We also indicate some attempts to find non-isomorphic graphs which are very similar in terms of the common subgraphs which they share.

In the following all graphs will be simple and undirected. Let $G$ be a graph and $v$, $e$ a vertex and an edge, respectively, of $G$. Then $G-v$ will denote the graph obtained by deleting from $G$ the vertex $v$ and all the edges incident to $v$; this will be called a *vertex-deleted subgraph* of $G$. More generally, if $X$ is a set of vertices of $G$ then $G-X$ will denote the graph obtained by deleting from $G$ all vertices in $X$ and all edges incident to at least one vertex in $X$. The resulting graph $G-X$ is said to be *induced* by the vertices $V(G)-X$.

Similarly, $G-e$ will denote the graph obtained by deleting the edge $e$; it will be called an *edge-deleted subgraph* of $G$. We shall mostly be concerned with vertex-deleted subgraphs, but we shall often indicate how the results and questions we present relate to the edge-deletion case.

The measure of similarity between two graphs which we shall be discussing is the number of vertex-deleted subgraphs which they possess in common. We define the *subgraph similarity* $\mathrm{sim}(G,H)$ between two graphs $G$ and $H$ with the same number of vertices $n$ as follows. Let $\mathcal{D}(G)$, called the *deck* of $G$, be the list of vertex-deleted subgraphs of $G$, where isomorphic subgraphs appear with the appropriate multiplicity. Similarly let $\mathcal{D}(H)$ be the deck of $H$. Then $\mathrm{sim}(G,H)$ is equal to the number of vertex-deleted subgraphs in $\mathcal{D}(G)$ which are also in $\mathcal{D}(H)$, where a subgraph which appears more than once in $\mathcal{D}(G)$ is counted as many times as it appears in $\mathcal{D}(H)$. (Therefore Zelinka's result quoted above for $k=1$ states that $G$ and $H$ are both in the deck of some graph if and only if $\mathrm{sim}(G,H)\geq 1$.) To make these definitions clear note that the two graphs $G$ and $H$ in Figure 1 have $\mathrm{sim}(G,H)=5$.

All of this is, of course, related to the Reconstruction Conjecture (RC) which can now be stated as:

**Reconstruction Conjecture**
*If $G$ and $H$ are two graphs on $n\geq 3$ vertices with $sim(G,H)=n$ then $G$ and $H$ are isomorphic.*

Most results in graph reconstruction can now be stated in this fashion, for example, if $\mathrm{sim}(G,H)=n$ then $G$ and $H$ have the same degree sequence, and

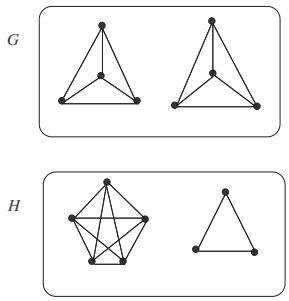


**Fig. 1.** Graphs $G$, $H$ with $\mathrm{sim}(G, H) = 5$

the same characteristic and chromatic polynomials [16, 17]; if $\mathrm{sim}(G, H) = n$ and $G$ is in one of these classes of graphs, then $G \simeq H$: regular, disconnected [17], trees [12] and maximal planar [15]. (For a survey on the RC the reader is referred to [16].)

But the new insight which this point of view brings is that now, new and perhaps more amenable structural questions about graphs arise. Basically, even if we assume that the RC is true, we can still ask questions such as how large can $\mathrm{sim}(G, H)$ be when $G$ and $H$ are not isomorphic. This enables us to revisit classes of graphs for which the question of reconstruction is easily settled but for which the issue of similarity in terms of subgraphs is still a very interesting unresolved question.

The notion of $\mathrm{sim}(G, H)$ is at the heart of two important parameters which have been studied in the literature on the RC. Both of these parameters indicate how similar or dissimilar a given graph is to all others, and therefore how easy or difficult it is to determine it from its deck. The *universal reconstruction number* $\forall\mathrm{rn}(G)$ of a graph $G$ is defined to be

$$\forall\mathrm{rn}(G) = 1 + \max_{H \not\simeq G}\{\mathrm{sim}(G, H)\}.$$

This means that, given *any* $\forall\mathrm{rn}(G)$ vertex-deleted subgraphs from $\mathcal{D}(G)$, these subgraphs determine $G$ uniquely because no other non-isomorphic graph can have all of them in its deck. This interpretation, which tacitly assumes that RC is true, explains the name given to this parameter and its notation. This parameter is also often called the *adversary reconstruction number* of $G$ [4].

The other reconstruction number $\exists\mathrm{rn}(G)$, called the *existential reconstruction number* of $G$, is defined a little differently. Again tacitly assuming the truth of the RC, $\exists\mathrm{rn}(G)$ is defined to be the smallest number of vertex-deleted subgraphs of $G$ which are not found in the deck of any other graph. This means that there exist $\exists\mathrm{rn}(G)$, and no less, vertex-deleted subgraphs of $G$ which alone determine $G$ uniquely, and this again explains the name of this parameter and the notation used. This parameter is also often called the *ally reconstruction number* or simply the *reconstruction number* of $G$ [4]. For reasons which will

become clear in the next section, in this chapter we shall mostly discuss these reconstruction numbers from the point of view of finding graphs with a high value for these parameters, that is, graphs which are in some sense very similar to other graphs. For more results about these two reconstruction numbers the reader is referred to the survey [4] and the book [17].

When we discuss the analogous situation with edge-deleted subgraphs we denote these parameters by the suffix $e$: $\mathrm{sim}_e$, $\forall\mathrm{rn}_e$ and $\exists\mathrm{rn}_e$.

It is clear that $\exists\mathrm{rn}(G) \leq \forall\mathrm{rn}(G)$ but sometimes the two can be equal. For the two graphs in Figure 1 one can check that $\exists\mathrm{rn}(G) = \forall\mathrm{rn}(G) = 6$, $\exists\mathrm{rn}(H) = 3$ and $\forall\mathrm{rn}(H) = 6$. It is also clear that $\exists\mathrm{rn}(G) > 2$ because suppose we claim that $\exists\mathrm{rn}(G) = 2$ for some graph $G$. Let $G - u$ and $G - v$ be the two vertex-deleted subgraphs which alone determine $G$. Construct $H$ as follows. If $u$ and $v$ are adjacent in $G$ then remove the edge $uv$, if they are not adjacent then add the new edge $uv$. Then, $H$ is not isomorphic to $G$ but it contains the two graphs $G - u$ and $G - v$ in its deck.

So the question becomes: how large can $\exists\mathrm{rn}(G)$ and $Urn(G)$ be? we have seen that for the graphs in Figure 1 $\exists\mathrm{rn}(G)$ is as small as it can be while $\forall\mathrm{rn}(G), \exists\mathrm{rn}(H)$ and $\forall\mathrm{rn}(H)$ are almost as large as the truth of the RC would allow. We shall see in the next section that such large reconstruction numbers are very rare.

## 2 Most graphs are dissimilar

It turns out that most graphs are so dissimilar that their universal reconstruction number is three, that is, any three vertex-deleted subgraphs of most graphs will determine its graph uniquely. We shall make this statement more precise and give the proof in full because it illustrates very well how the concept of subgraph similarity which we are using depends heavily on the internal structure of graphs. The proof is based on [1] (Chapter 10, "Probabilistic Lens: Counting subgraphs").

It is well known that almost every graph has a trivial automorphism group . However, a stronger results is possible which will tell us a lot about $\mathrm{sim}(G, H)$, but we need first to explain what we mean when we say that almost every graph has some property. So, let $\mathcal{P}$ be a graph theoretic property such as 'planar' or 'vertex-transitive'. Let $r_n$ denote the proportion of labelled graphs on $n$ vertices that have property $\mathcal{P}$. If $\lim_{n\to\infty} r_n = 1$, then we say that *almost every (a.e.) graph has property* $\mathcal{P}$. To show that a.e. graph has our desired property we will use the simplest probability space which is set up when studying random graphs. Let $\mathcal{G}(n, \frac{1}{2})$ be the set of all labelled graphs on the set of vertices $\{1, 2, \ldots, n\}$ where, for each pair $i, j$,

$$\mathrm{P}(ij \text{ is an edge}) = \mathrm{P}(ij \text{ is not an edge}) = \frac{1}{2}$$

independently. Therefore each graph $G$ in $\mathcal{G}(n, \frac{1}{2})$ has probability $(\frac{1}{2})^{\binom{n}{2}}$, which is, of course, equal to the probability of choosing $G$ randomly from amongst all $2^{\binom{n}{2}}$ labelled graphs on $n$ vertices when all are equally likely to be chosen. So, in order to show that a.e. graph has a particular property $\mathcal{P}$ one has to show that the probability that $G \in \mathcal{G}(n, \frac{1}{2})$ has property $\mathcal{P}$ tends to 1 as $n$ tends to infinity.

The property we are interested in is the following. Let $k$ be fixed. We say that a graph $G$ has *property* $A_k$ if all induced subgraphs of $G$ on $n-k$ vertices are mutually nonisomorphic. In other words, $G$ has property $A_k$ means that, if $X, Y$ are two distinct $k$-subsets of $V(G)$, then $G - X \not\cong G - Y$. It is easy to see that if $G$ has property $A_{k+1}$, then it also has property $A_k$ and that if it has property $A_1$, then it is asymmetric. Therefore having property $A_k$ is stronger than just being asymmetric. We shall show that, for any fixed $k$, a.e. graph has property $A_k$.

**Lemma 1.** *Let $W \subseteq V$, $|W| = t$, $|V| = n$, and let $\rho : W \to V$ be an injective function that is not the identity. Let $g = g(\rho)$ be the number of elements $w \in W$ such that $\rho(w) \neq w$. Then there is a set $I_\rho$ of pairs of (distinct) elements of $W$, containing at least $2g(t-2)/6$ pairs, such that $I_\rho \cap \rho(I_\rho) = \emptyset$.*

Consider those pairs $v, w \in W$ such that at least one is moved. (All pairs are taken to contain distinct elements.) There are $g(t-g) + \binom{g}{2}$ such pairs. For all but at most $g/2$ of these pairs, $\{v, w\} \neq \{\rho(v), \rho(w)\}$ (the exceptions are when $\rho(v) = w$ and $\rho(w) = v$). Let $E_\rho$ be the set of all such pairs. Then

$$|E_\rho| \geq g(t-g) + \binom{g}{2} - \frac{g}{2} = g(t - \frac{g}{2} - 1) \geq g(\frac{t}{2} - 1).$$

Define a graph $H_\rho$ with vertex-set the pairs in $E_\rho$ and such that each pair $\{v, w\}$ is adjacent to the pair $\{\rho(v), \rho(w)\}$. In $H_\rho$, all degrees are at most 2. Degrees equal to 1 could arise because $\{\rho(v), \rho(w)\}$ could contain an element not in $W$, and so the pair would not be in $E_\rho$. Degrees equal to 2 could arise because $\{v, w\}$ could be adjacent to both $\{\rho(v), \rho(w)\}$ and $\{\rho^{-1}(v), \rho^{-1}(w)\}$.

Therefore the components of $H_\rho$ are isolated vertices, paths or cycles. Let $I_\rho$ be a set of independent (that is, not adjacent) vertices in $H_\rho$. Therefore, for any pair $\{v, w\} \in I_\rho$, $\{\rho(v), \rho(w)\}$ is not in $I_\rho$.

Now, all isolated vertices in $H_\rho$ are independent, at least half of the vertices on a path are independent and at at least one third of the vertices on a cycle are independent, the extreme case here being a triangle. Therefore

$$|I_\rho| \geq |E_\rho|/3 \geq \frac{2g(t-2)}{6},$$

as required. □

**Corollary 1** *Let $G \in \mathcal{G}(n, \frac{1}{2})$, $W \subset V = V(G)$ and $|W| = t$. Let $\rho : W \to V$ be an injective function that is not the identity. Let $g = g(\rho)$ be the number of elements $w \in W$ such that $\rho(w) \neq w$. Let $S_\rho$ be the event*

*"$\rho$ gives an isomorphism from $G[W]$ to $G[\rho(W)]$".*

*Then*

$$P(S_\rho) \leq \left(\frac{1}{2}\right)^{2g(t-2)/6}.$$

Let $I_\rho$ be the set constructed in the previous lemma. Now, for a given pair $\{v, w\} \in I_\rho$, the event

"$\{v, w\}$ and $\{\rho(v), \rho(w)\}$ are both edges or nonedges"

has probability $1/2$. These events, as they range over all pairs $\{v, w\} \in I_\rho$, are mutually independent, because they involve distinct pairs. But $S_\rho$ requires all these events simultaneously. Therefore, by independence,

$$P(S_\rho) \leq (\frac{1}{2})^{|I_\rho|} \leq \left(\frac{1}{2}\right)^{2g(t-2)/6},$$

as required. □

The result of this corollary is the crux of the matter. There are too many independent correct 'hits' required for $\rho$ to be an isomorphism, and the probability therefore becomes small as $n$ increases.

**Theorem 2 (Korshunov [14]; Müller [21]; Bollobás [6]).** *Let $k$ be a* fixed *nonnegative integer and let $G \in \mathcal{G}(n, \frac{1}{2})$. Let $p_n$ denote the probability that*

$$\exists W \subseteq V(G) = V = \{1, 2, \ldots, n\},$$

*with $|W| = n - k$ and such that*

$$\exists \rho : W \to V, \rho \neq id, \rho \text{ is an isomorphism from } G[W] \text{ to } G[\rho(W)].$$

*Then,* $\lim_{n\to\infty} p_n = 0$.

*Hence, a.e. graph has property $A_k$.*

Pick a particular $W \subset V$ with $|W| = n - k$. This can be done in $\binom{n}{n-k}$ ways, and

$$\binom{n}{n-k} = \frac{n(n-1)\ldots(n-k+1)}{k!} < n^k.$$

Let $t = n - k$. Let $\rho : W \to V$ be injective and not the identity, and let $g = g(\rho)$ be the number of vertices of $W$ that are moved by $\rho$. Let $S_\rho$ be the event defined in the previous corollary.

Now, for a given value of $g$ between 1 and $t$, how many functions $\rho$ are there such that $g(\rho) = g$? Such a function is determined by the set $\{w : \rho(w) \neq w\}$ and by the values it takes on this set. Therefore, there are less than $n^{2g}$ such $\rho$. Therefore, for a given fixed $W$, the probability of a nontrivial isomorphism is given by

$$\begin{aligned}\sum_{\rho \neq \mathrm{id}} P(S_\rho) &= \sum_{g=1}^{t} \sum_{\rho : g(\rho) = g} P(S_\rho) \\ &\leq \sum_{g=1}^{t} n^{2g} \left(\frac{1}{2}\right)^{2g(t-2)/6} \\ &= \sum_{g=1}^{t} \left[n^2 2^{(2-t)/3}\right]^g \\ &< \sum_{g=1}^{t} \left[4^{1/3} n^2 2^{-t/3}\right]^g .\end{aligned}$$

Now $t = n - k > 12(k+1)\lg n$ for sufficiently large $n$. Therefore

$$\begin{aligned}4^{1/3} n^2 2^{-t/3} &< 4^{1/3} n^2 2^{-4(k+1)\lg n} \\ &= \frac{4^{1/3} n^2}{n^{4(k+1)}} \\ &\leq \frac{4^{1/3}}{n^{2(k+1)}} \\ &< \frac{1}{n^{k+1}}\end{aligned}$$

where the last inequality follows if $4^{1/3} < n^{k+1}$.

Therefore

$$\begin{aligned}\sum_{\rho \neq \mathrm{id}} P(S_\rho) &< \sum_{g=1}^{t} \left(\frac{1}{n^{k+1}}\right)^g \\ &< \sum_{g=1}^{n} \left(\frac{1}{n^{k+1}}\right)^g \\ &= \frac{n^{n(k+1)} - 1}{n^{n(k+1)}(n^{k+1} - 1)}.\end{aligned}$$

But all this is for fixed $W$. Therefore the required probability is

$$p_n < n^k \frac{n^{n(k+1)} - 1}{n^{n(k+1)}(n^{k+1} - 1)},$$

and this tends to 0 as $n$ tends to infinity. □

Now the following theorem explains the relationship between property $A_k$ and the subgraph similarity between graphs which we have been discussing.

**Theorem 3 (Müller [21]; Myrvold [22]; Bollobás [6]).** *Let $G$ have property $A_3$. Then $G$ can be uniquely determined from any three vertex-deleted subgraphs in its deck. That is, $sim(G, H) \leq 2$ for any graph $H$ not isomorphic to $G$ and $\forall rn(G) = 3$.*

Let $u, v, w \in V(G)$. We shall show that $G$ is uniquely determined from just $G - u$, $G - v$ and $G - w$.

Note first that $v$ is identifiable in $G - u$ and $u$ is identifiable in $G - v$; because, since $G$ has property $A_3$ (and hence $A_2$), the only pair of vertices $x \in V(G - u)$, $y \in V(G - v)$ such that $G - u - x \simeq G - v - y$ are $x = v$ and $y = u$. Let $X = G - u - x$ and $Y = G - v - y$. There can only be one isomorphism from $X$ to $Y$. For suppose $\alpha$ and $\beta$ are two such isomorphisms. Let $z \in V(X)$ such that $\alpha(z) \neq \beta(z)$. Then $X - z \simeq Y - \alpha(z) \simeq Y - \beta(z)$, contradicting property $A_3$. Therefore we can label $X$ and $Y$ uniquely, and, from $X = G - u$, we can determine uniquely all of the neighbours of $v$ in $G$, except possibly $u$. All we need to know is whether $u$ and $v$ are adjacent. To determine this we repeat the above procedure with $G - w$ instead of $G - u$. □

From Theorem 2 and this lemma the following surprising result is immediate.

**Theorem 4.** *Almost every graph $G$ has $sim(G, H) \leq 2$ for any graph $H \not\simeq G$ and therefore $\forall rn(G) = 3$.*

In an analogous manner one prove this result on edge-deleted subgraphs.

**Theorem 5.** *Almost every graph $G$ has the property that any two edge-deleted subgraphs from its edge-deck determine it uniquely, that is, $sim_e(G, H) \leq 1$ for any graph $H \not\simeq G$, and $\forall rn_e(G) = 2$.*

### 2.1 Empirical evidence

The data in Table 2.1, obtained by McMullen and Radziszowski [18], gives a very good idea of how strong Theorem 4 really is. Out of more than 12,000,000, graphs on ten vertices, only twelve have $\exists$rn greater than the minimum possible value of 3. This situation sets the scene for the search of graphs with large values of $\exists$rn and $\forall$rn, and sometimes even a value of four can be considered large and graphs with this value could be difficult to find. In the next section we shall look at some results which have been obtained in this vein.

## 3 Graphs with large subgraph similarity

We shall look at the problem of finding graphs with large subgraph similarity from two angles, that of the existential reconstruction number $\exists$rn and the universal reconstruction number $\forall$rn.

**Table 1.** *Number of graphs with given order and given ∃rn*

| ∃rn | Order | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 3 | 4 | 8 | 34 | 150 | 1044 | 12, 334 | 274, 666 | 12, 005, 156 |
| 4 | | 3 | | 4 | | 8 | | 6 |
| 5 | | | | 2 | | 2 | 2 | 4 |
| 6 | | | | | | 2 | | |
| 7 | | | | | | | | 2 |

### 3.1 Large values of ∃rn

The first graphs for which ∃rn were studied were disconnected graphs. Myrvold [23] and Molina [20] showed the following.

**Theorem 6.** *A disconnected graph with non-isomorphic components has ∃rn equal to 3. A disconnected graph with all components isomorphic each having c vertices has ∃rn* $\leq c + 2$.

So here it seems that we have a rich supply of graphs with large ∃rn. The example which Myrvold gave of disconnected graphs with ∃rn $= c+2$ was the graph $G$ consisting of disjoint copies of the complete graph $K_c$. The graph $G$ in Figure 1 is the special case $K_4 \cup K_4$. However, Asciak and Lauri [5] showed that in fact these are the only examples of disconnected graphs with ∃rn $= c + 2$ and that there are no disconnected graphs with ∃rn $= c + 1$. The computer searches of McMullen and Radziszowski [18] amongst all graphs on at most ten vertices unearthed only two examples of disconnected graphs with ∃rn $> 3$. These are the graph made up of two disjoint copies of the cycle on four vertices and the graph made up of two disjoint copies of the path on four vertices. Both have ∃rn $= 4$ and no other disconnected graphs with ∃rn $> 3$ are known. The big gap between ∃rn $= 4$ and ∃rn $= c$ is waiting to be explored.

The situation with regular graphs is somewhat similar. Myvold [22] has shown that $r$-regular graphs have ∃rn at most $r + 3$ but Asciak [3] has shown that again the disconnected graph consisting of disjoint copies of $K_{r+1}$ is the only $r$-regular graph with ∃rn $= r + 3$. Here too, knowledge about the gap between ∃rn $= 4$ and ∃rn $= r + 2$ is very scant. The computer searches of McMullen and Radziszowski led them to this construction. The graph $RCC_{n,j}$ is obtained as follows. Take $n \geq 2$ disjoint copies of cycles each of length $j \geq 3$. Let $v_{c,i}, i \in \{0, 2, \ldots, j-1\}$ denote the $i$-th vertex of the $c$-th cycle. For each $c \neq d$ join the vertices $v_{c,i}$ and $v_{d,i+1}$, where addition is modulo $j$. The resulting graph $RCC_{n,j}$ is regular and McMullen and Radziszowski prove the following.

**Theorem 7.** *$\exists rn(RCC_{n,j}) > n+1$, for all $n \geq 2$ and $j \geq 3$.*

However, no other regular graphs with $\exists\text{rn} > 3$ are known and, as McMullen and Radziszowski say, there seems to be no clear idea on how to establish in general the exact value of $\exists\text{rn}(RCC_{n,j})$ for all $n, j$.

These cases illustrate the new set of problems which the notion of reconstruction numbers creates. The classical reconstruction of regular graphs is trivial and that of disconnected graphs is an easy exercise [17]. But even finding examples with $\exists\text{rn} > 3$ is a difficult task. The reader who is interested in finding out more about graphs with large $\exists$rn is invited to read [18].

### 3.2 Large values of $\forall$rn

The definition of $\forall$rn is more closely related to that of $\text{sim}(G, H)$, and it seems more difficult to tackle. It certainly seems easier to find disconnected graphs with large $\forall$rn than ones with large $\exists$rn. For example, Hemaspaandra et al [11] observe that since

$$\text{sim}(K_{t+1} \cup K_{t-1}, 2K_t) = t + 1$$

then $\forall\text{rn}(K_{t+1} \cup K_{t-1}$ and $\forall\text{rn}(2K_t)$ are both at least $t + 2$ and therefore greater than the corresponding $\exists$rn numbers which are both three. However, the proof in [11] that these two $\forall$rn numbers are actually $t + 2$ is not simple, even for such straightforward graphs, showing that determining $\forall$rn seems to be quite difficult in general. Also, it is not clear that these and the other two examples given in [11] are not exceptional cases similar to the usual suspects: the graphs $pK_n$ with large $\exists$rn. Therefore the question of finding disconnected graphs with large $\forall$rn might be as open as it is for finding disconnected or regular graphs with large $\exists$rn.

Until recently, most of the results obtained about $\forall$rn and $\text{sim}(G, H)$ were found in [22] and [10]. An early result was the following.

**Theorem 8 (Myrvold [22]).** *Let $G$ and $H$ be two graphs on $n$ vertices and with $sim(G, H) = n - 1$. Then $G$ and $H$ have the same degree sequence.*

Again we see that what is an easy exercise in reconstruction [17] becomes a difficult result when seen in terms of the subgraph similarity between graphs. The obvious, and difficult, question here is: for given $n$, what is the largest value of $k$ such that there exist graphs $G$ and $H$ on $n$ vertices with $\text{sim}(G, H) = k$ but with different degree sequences.

Of course, the most general problem here is to determine the largest value of $\text{sim}(G, H)$ for non-isomorphic graphs on $n$ vertices. But since this would solve the RC, all authors have attempted this question by restricting $G$ and $H$ to particular classes and generally trying to determine the maximum possible value of $\text{sim}(G, H)$.

Significant advances in this direction have recently been reported by Bowler, Brown and Fenner in [7]. For example, they show the following.

**Theorem 9.** *Let $G$ be a tree and $H$ a unicyclic graph on $n$ vertices ($n \geq 19$). Then*

$$sim(G, H) \leq \lfloor \frac{2}{5}(n+1) \rfloor.$$

*Moreover, this bound is attained.*

From this result and other work in [22] the following holds.

**Theorem 10.** *Let $G$ and $H$ be two graphs on $n$ vertices ($n \geq 19$) and such that*

$$sim(G, H) \geq \lfloor \frac{n}{2} \rfloor + 1.$$

*Then if $G$ is a tree $H$ must also be a tree.*

Francalanza [10] also considered the number of edge-deleted subgraphs in common between a tree and a unicyclic graph plus an isolated vertex. She proved the following.

**Theorem 11.** *Let $G$ be a tree and $H$ a unicyclic graph with an isolated vertex, both on $n$ vertices. Then*

$$sim_e(G, H) \leq \frac{n}{2} + 1.$$

Bowler, Brown and Fenner make conjecture that if $G$ is a tree and $H$ is a unicyclic graph plus an isolated vertex, both on $n$ vertices, then in fact

$$\mathrm{sim}_e(G, H) \leq \frac{n}{2}.$$

The structure of the trees and unicyclic graphs which attain large subgraph similarity between them have are very particular. The trees are *caterpillars*, that is, trees the deletion of whose endvertices gives a path, and the unicyclic graphs are what Myrvold and Francalanza call *sunshine graphs*, that is, unicyclic graphs the deletion of whose endvertices gives a cycle.

The main question which these researchers would like to answer here is certainly the following: What is the largest possible value of $\mathrm{sim}(G, H)$ when $G$ and $H$ are two non-isomorphic trees on $n$ vertices?

This construction from [7] gives a family of pairs of non-isomorphic trees with large subgraph similarity. Let

$$G^* = K_{1,p-1} \cup K_{1,p+1} \cup K_{1,p+1}$$
$$H^* = K_{1,p} \cup K_{1,p} \cup K_{1,p+1}.$$

Let $G$ be the tree obtained from $G^*$ by adding a new central vertex and three new edges joining the new vertex to the three cutvertices of $G^*$. Similarly, construct $H$ from $H^*$. These two trees are non-isomorphic, have $n = 3p + 5$ vertices, and

$$\mathrm{sim}(G,H) = 2p = \frac{2}{3}(n-5).$$

This family of tree pairs has the highest known subgraph similarity between non-isomorphic trees. A similar construction in [7] gives examples of pairs $G, H$ of non-isomorphic trees on $n$ vertices with the same degree sequence and

$$\mathrm{sim}(G,H) = \frac{2}{3}(n+1-2\sqrt{3n-6}).$$

The best result known to date regarding the highest possible value of $\mathrm{sim}(G,H)$ for general graphs is again given found in [7]. First we require a definition. A *2UC graph pair* is a pair of non-isomorphic graphs, $G$ and $H$, on $n$ vertices, at least one of which is disconnected, such that in $G$ or in $H$ there are at least two components which cannot be matched with the components of the other graph by isomorphism. A particular example is when $G$ is connected and $H$ is disconnected. ("2UC" stands for "Two Unmatched Components".) The motivation behind this definition is that if $A$ and $B$ are two non-isomorphic connected graphs with the same deck (hence counterexamples to the RC) and on $n-1$ vertices, then $\mathrm{sim}(A \cup K_1, A \cup K_1) = n-1$. Bowler, Brown and Fenner prove the following theorem.

**Theorem 12.** *Let $G$ and $H$ be two 2UC graphs. Then*

$$sim(G,H) \leq 2\lfloor \frac{1}{3}(n-1) \rfloor.$$

For $n \geq 22$ and $n \equiv 1 (\mathrm{mod} 3)$, they also give the following infinite family of pairs of 2UC graphs attaining this bound:

$$G = K_{p-1} \cup K_{p+1} \cup K_{p+1}$$
$$H = K_p \cup K_p \cup K_{p+1}.$$

They also show that this pair is unique for the given values of the parameter $n$. Note that although $G$ and $H$ are disconnected, their complements are connected and also have the same subgraph similarity.

More examples are given in [7] including uniqueness of some families of pairs attaining the upper bound in Theorem 12. Their work also gives an example of pairs $G, H$ of 2UC graphs with $n = 3p^2 - 2, (p \geq 3)$, having the same degree sequence, and

$$\mathrm{sim}(G,H) = \frac{2}{3}(n+5-2\sqrt{3n+6}).$$

This number is smaller than the upper bound in Theorem 12. Therefore it seems natural to ask what is the maximum possible value of $\mathrm{sim}(G,H)$ when $G, H$ are two non-isomorphic 2UC graphs on $n$ vertices with the same degree sequence.

Motivated by Theorem 12, Bowler, Brown and Fenner make the following conjecture which, of course, is a considerable strengthening of the RC.

**Strong Reconstruction Conjecture**
*Let $G$ and $H$ be non-isomorphic graphs on $n$ vertices. For large enough $n$,*

$$sim(G, H) \leq 2\lfloor \frac{1}{3}(n-1) \rfloor.$$

*Therefore for any graph $G$ on $n$ vertices and sufficiently large $n$,*

$$\forall rn(G) \leq 2\lfloor \frac{1}{3}(n-1) \rfloor + 1.$$

Finally, what about $\forall \mathrm{rn}_e(G)$, the universal edge-reconstruction number. In classical graph reconstruction, determining $G$ from edge-deleted subgraphs is always easier than determining it from vertex-deleted subgraphs. However, the relationship between the vertex and the edge versions of the parameters which we have been discussing in this chapter does not seem to be so straightforward (see [4] for more on this). Sometimes the edge parameter is larger than the corresponding vertex parameter, and often determining the former is at least as difficult as finding the latter. Certainly, very little work, if any, has been done on $\forall \mathrm{rn}_e(G)$, especially the search for graphs with large $\forall \mathrm{rn}_e$, so this is a field wide open for investigation.

## 4 Algorithmic and other issues

The RC is not an algorithm question. The issue is not whether there is an efficient way of obtaining $G$ from its deck but it is a question of uniqueness: is there more than one graph with the given deck? However, a few variants of the RC have been adapted into a question of algorithmic complexity. Subgraph similarity and reconstruction numbers, being so closely related Graph Isomorphism Problem (GI), and the Subgraph Isomorphism Problem which is known to be NP-complete [13] are perhaps the most natural variants of the reconstruction problem to be treated algorithmically.

In [11], the authors define these four decision problems.

1. EXIST-VRN $= \{\langle G, k\rangle | \exists \mathrm{rn}(G) \leq k\}$.
2. UNIV-VRN $= \{\langle G, k\rangle | \forall \mathrm{rn}(G) \leq k\}$.
3. EXIST-ERN $= \{\langle G, k\rangle | \exists \mathrm{rn}_e(G) \leq k\}$.
4. UNIV-ERN $= \{\langle G, k\rangle | \forall \mathrm{rn}_e(G) \leq k\}$.

They remark that it is easy to see that EXIST-VRN $\in \Sigma_2^p$ (since GI is low for $\Sigma_2^p$), UNIV-RN $\in \mathrm{coNP}^{\mathrm{GI}}$, EXIST-ERN $\in \mathrm{NP}^{\mathrm{GI}}$ and UNIV-ERN $\in \mathrm{coNP}^{\mathrm{GI}}$ and they suggest that obtaining tight, or tighter, bounds on the

complexity of these problems should be interesting. (For explanations of the above complexity terms the reader is referred to [13].)

And finally, what about possible applications. Any measure of similarity between mathematical objects is bound to have some relevance in situations modeled by the objects, and graphs are certainly amongst the mathematical structures most often used as models. One in which notions which related to the concept of subgraph similarity seem to be useful is in systems biology. The way a cell processes information from its environment in order to determine the rate of production of the proteins it requires is often modeled by what are called *transcription networks*, which are basically directed graphs [2]. Biologists try to identify particular subgraphs of transcription networks in order to explain their functionality. These *network motifs* are often identified as those subgraphs in the transcription network which appear significantly more often than they do in a random graph of the same size. This seems quite reminiscent of the notion we have been discussing of comparing two graphs by counting the number of subgraphs they have in common. Here, the comparison is usually between the given transcription network and the general random graph. In this comparison, the number of symmetries of the network motif (the size of its automorphism group) often plays an important part. The notion of how a subgraph embeds in a graph, a notion which involves the number of appearances of the subgraph and the size of its automorphism group, seems to be the central issue in the reconstruction problem (see, for example, Chapter 10 and especially Chapter 11 in [17]. An investigation of how these ideas from subgraph similarity and graph reconstruction might apply to the study of network motifs in transcription networks could therefore be very useful

Similar ideas have cropped up in the unlikely area of counter-terrorism! Interactions between agents in a society (conversations, emails, telephone calls, etc) can be modeled by a graph. Within this "transactional noise" one would like to detect the emergence of unlikely configurations (subgraphs) which could signify the existence of networks of terrorist activities [19]. Again, this is done by comparing the transactional network with some appropriate random graph model to detect subgraphs which appear more frequently than expected by the model. The similarities with the previous application and what we have been discussing is clear.

It is, after all, not surprising that such applications should exist. With the ability to collect and handle ever larger amounts of data in various fields from biology to sociology comes the need of modeling situations with large graphs. And very often a natural way to investigate certain aspects of the internal structure of such graphs is through smaller subgraphs which are more manageable. And these applications are often closely related to issues of algorithmic complexity. When tackling empirically questions about subgraphs in common between two graphs one cannot escape the from the Graph Isomorphism Problem in some guise or another.

## 5 Conclusion

We have discussed a way of measuring similarity between graphs in terms of subgraphs which does not simply give an alternative framework for wording the Reconstruction Conjecture. It raises simple questions which are difficult to solve about graphs which are very easily reconstructible, and it gives some new twists to old ideas, such as the relationship between vertex-reconstruction and edge-reconstruction. Independently of the status of RC, finding classes of graphs with large subgraph similarity or reconstruction number is an interesting non-trivial problem. And the notion of comparing graphs in terms of the number of common subgraphs of some type or another that they share seems to be a promising area of modern applied graphs theory, which is closely connected to algorithmic complexity issues related to reconstruction numbers, which are in turn of important theoretical interest. It seems that subgraph similarity has a lot to offer to graph theorists with different interests and tastes.

## References


1. N. Alon and J.H. Spencer. *The Probabilistic Method*. Wiley, 1992.
2. U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall CRC Press, 2007.
3. K.J. Asciak. On certain classes of graphs with large reconstruction number. Master's thesis, University of Malta, 1998.
4. K.J. Asciak, M.A. Francalanza, and J. Lauri W. Myrvold. A survey of some open questions in reconstruction numbers.
5. K.J. Asciak and J. Lauri. On disconnected graphs with large reconstruction number. *Ars Combinatoria*, 62:173–181, 2002.
6. B. Bollobás. Almost every graph has reconstruction number 3. *J. Graph Theory*, 14:1–4, 1990.
7. A. Bowler, P. Brown, and T. Fenner. Families of pairs of graphs with a large number of common cards. *Preprint*.
8. H. Bunke. Recent developments in graph matching. In *Proceedings of the 15th Int. Conf. on Pattern Recognition, Spain*, volume 2, 2000.
9. M. Dehmer and A. Mehler. A new method of measuring similarity for a special class of directed graphs. *Tatra Mt. Math. Publ.*, 36:39–59, 2007.
10. M. A. Francalanza. The adversary reconstruction of trees: The case of caterpillars and sunshine graphs. Master's thesis, University of Malta, 1999.
11. Edith Hemaspaandra, Lane A. Hemaspaandra, Stanisław P. Radziszowski, and Rahul Tripathi. Complexity results in graph reconstruction. *Discrete Appl. Math.*, 155:103–118, 2007.
12. P. J. Kelly. A congruence theorem for trees. *Pacific J. Math.*, 7:961–968, 1957.
13. J. Köbler, U. Schöning, and J. Torán. *The Graph Isomorphism Problem: Its Structural Complexity*. Birkhäuser, 1993.
14. A. D. Korshunov. Number of nonisomorphic graphs in an $n$-point graph. *Math. Notes of the Acad. USSR*, 9:155–160, 1971.

15. J. Lauri. The reconstruction of maximal planar graphs, II: Reconstruction. *J. Combin. Theory (Ser. B.)*, 30:196–214, 1981.
16. J. Lauri. The Reconstruction Problem. In J. L. Gross and J. Yellen, editors, *Handbook of Graph Theory*, pages 79–98. CRC Press, 2004.
17. J. Lauri and R. Scapellato. *Topics in Graph Automorphisms and Reconstruction*. Cambridge University Press, 2003.
18. B. McMullen and S.P. Radziszowski. Graph reconstruction numbers. *J. Combin. Math. Combin. Comp.*, 62:85–96, 2005.
19. T. Mifflin, C. Boner, G. Godfrey, and M. Greenblatt. Detecting terrorist activities in the twenty-first century: A theory of detection for transactional networks. In R.L. Popp and J. Yen, editors, *Emergent Information Technologies and Enabling Policies for Counter-Terrorism*, Series on Computational Intelligence, pages 349–365. IEEE Press, 2006.
20. R. Molina. Correction of a proof on the ally-reconstruction number of a disconnected graph. *Ars Combinatoria*, 40:59–64, 1995.
21. V. Müller. The edge reconstruction hypothesis is true for graphs with more than $n \log_2 n$ edges. *J. Combin. Theory (Ser. B)*, 1975?
22. W. Myrvold. *Ally and Adversary Reconstruction Problems*. PhD thesis, University of Waterloo, Ontario, Canada, 1988.
23. W. J. Myrvold. The ally-reconstruction number of a disconnected graph. *Ars Combin.*, 28:123–127, 1989.
24. R.C. Read and D.G. Corneil. The graph ismorphism disease. *J. Graph Theory*, 1:339–363, 1977.
25. B. Zelinka. On a certain distance between isomorphism classes of graphs. *Časopis pro pěst. Mathematiky*, 100:371–373, 1975.

# Index


$G - X$, 2
$G - e$, 2
$G - v$, 2
$\mathcal{D}(G)$, 2
$\exists\text{rn}(G)$, 3
$\forall\text{rn}(G)$, 3
$\text{sim}(G, H)$, 2
$r$-regular graphs, 9
$\mathcal{G}(n, \frac{1}{2})$, 4

a.e., 4
adversary reconstruction number, 3
ally reconstruction number, 3
almost every, 4
automorphism group, 4

caterpillar, 11
characteristic polynomials, 3
chromatic polynomials, 3

deck, 2
degree sequence, 2, 10
distance between graphs, 1

edge-deleted subgraph, 2
existential reconstruction number, 3

GI, 1, 13
Graph Isomorphism Problem, 1, 13

induced subgraph, 2
inexact graph matching, 1

network motifs, 14

property $A_k$, 5

RC, 2
Reconstruction Conjecture, 2
reconstruction number, 3

similarity between two graphs, 2
Subgraph Isomorphism Problem, 13
subgraph similarity, 2
sunshine graph, 11
systems biology, 14

transcription networks, 14

unicyclic graphs, 11
universal reconstruction number, 3

vertex-deleted subgraph, 2