Some Effects of the Human Genome Project on the Erdős Collaboration Graph

Chris Fields

Sonoma, CA 95476, USA
fieldsres@gmail.com

Abstract

The Human Genome Project introduced large-scale collaborations involving dozens to hundreds of scientists into biology. It also created a pressing need to solve discrete mathematics problems involving tens of thousands of elements. In this paper, we use minimal path lengths in the Erdős Collaboration Graph between prominent individual researchers as a measure of the distance between disciplines, and we show that the Human Genome Project brought laboratory biology as a whole closer to mathematics. We also define a novel graph reduction method and a metric that emphasizes the robustness of collaborative connections between researchers; these can facilitate the analysis of both within- and between-community connectivity in collaboration graphs.

1. Introduction

The structure of the Erdős Collaboration Graph, the graph of researchers (vertices) and joint publications (edges) that contains Paul Erdős as a vertex, provides a fascinating window into the relationships between academic disciplines, and in particular into the relationships between mathematics and the various sciences. The Erdős Number Project at Oakland University, for example, lists Erdős numbers - minimum path lengths from a researcher to Erdős in the Erdős Collaboration Graph - for Nobel Prize winners in physics starting from 1914, Chemistry from 1936, Medicine from 1958 and Economics from 1970.[1] Given the prominent roles typically played by Nobel

[1]Data files and other information are available at http://www.oakland.edu/enp/, accessed on June 24, 2014.

Prize winners in their respective research communities, the temporal depth of these lists suggests that the collaboration graph spanning all of the sciences has had a single, multi-disciplinary giant component for many years, and that this giant component is the Erdős Collaboration Graph. It is natural, moreover, to expect that individual disciplines and even subdisciplines form distinctive clusters within this giant component. Consistent with this expectation, Newman [29] showed that subdiscipline-specific collaboration graphs have distinctive community structures, and in particular that collaborations in the biomedical sciences exhibited far less clustering - and hence greater apparent randomness - than collaborations in subdisciplines of physics or the information sciences. Distinctive community structures and disciplinary and subdisciplinary clustering have been confirmed by subsequent studies of coauthorship in discipline-specific data sets [31] and of citation patterns in multi-disciplinary data sets [22, 28, 40]. Such community structure is also observed in metabolic, electronic, and other non-social networks (e.g. [21]).

The existence of community structure in coauthorship networks leads naturally to the question of how distinct communities are connected. One can ask, for example, whether connections between communities involve researchers who are peripheral or central, by some appropriate definition, to their primary communities, and whether communities are joined by one or a few "weak links" or by robust subgraphs. These questions can be formulated in terms of the notion of "betweenness centrality," a measure of the number of paths between disparate researchers that traverse a given researcher [6, 14] and the closely-related functional concept of "brokerage" of information passing between two nodes in a network by some intermediate node that the information-transfer path traverses [7, 23]. In these terms, one can ask whether brokers of information and hence collaboration between communities tend to be peripheral or central within their primary communities, and whether there tend to be few or many information brokers between pairs of identified communities.

The present paper considers these questions of betweenness and brokerage by examining the effects of a particular episode in the history of biology, the Human Genome Project (HGP), on the local fine structure and consequentially on the larger-scale community structure of the Erdős Collaboration Graph. The initial genomic DNA sequencing component of the HGP was completed in 2001 [25, 46]; for historical overviews, see [37, 50]. The HGP introduced to molecular biology the kind of massive multi-institutional

collaborations that had previously only characterized some areas of experimental physics. The HGP also introduced, even in its early, pre-sequencing stages, a data analysis problem new to biology: the problem of assembling coherent networks of entities from noisy, error-ridden data on the relationships between pairs of entities. The challenges posed by this novel network assembly problem, and by the increasing demands for DNA sequence analysis and laboratory data management, led to the emergence of the new interdisciplinary field of bioinformatics. As Grossman in [17] noted, two prominent participants in the HGP, E. Koonin and E. Lander, have Erdős numbers of 2, and hence many participants in the HGP have finite Erdős numbers due to research collaborations with these two leaders of the bioinformatics movement. Indeed, "it is probably possible to connect to Erdős a large fraction of people who have published in the biological sciences" [17, page 41] due to collaborative links leading back to the HGP.

The increasing size of highly-visible collaborative teams throughout the sciences has been widely noted (e.g. [5]), and one might expect the existence of such collaborations to have clearly-ascertainable effects on the large-scale structure of the Erdős Collaboration Graph. However, large-scale studies of several research communities over the relevant time frame have revealed only modest increases in the average numbers of authors of published papers, a rough guide to average collaborative team size. Newman [31], for example, reports an average of 3.75 authors per paper in the biomedical sciences between 1995 and 1999 (inclusive); Porter and Rafols [36] report averages of 4.9 authors per paper in medical science and 6.1 in neuroscience in 2005, while Wallace, Larivière and Gingras [47] report averages of 5.1 authors per paper in biochemistry and molecular biology and 4.9 in neuroscience in 2006. The impacts of large collaborations on average measures may, therefore, be much smaller than their scientific or technological significance might suggest. The studies reported in [36, 47] reveal, moreover, only modest increases in citation-based, average measures of interdisciplinarity in both biological and other disciplines over the past three decades. Global "map of science" measures show, in particular, a continuing separation between the large, dense cluster of laboratory-based biological sciences and even "interdisciplinary" mathematics [36, 28]. Investigating the effects of the HGP on the relationship between laboratory biology and mathematics, and on the structure of the Erdős Collaboration Graph in particular, would therefore appear to require a more focused and sensitive approach.

While scientific disciplines and subdisciplines are often defined for the purposes of large-scale studies by collections of scientific journals, the HGP can at best be loosely defined as a collection of research groups, many of which were assembled *ad hoc*, that were funded by public or private entities under the HGP rubric. Therefore a historical approach centered on the leaders of these groups provides an alternative, even if potentially biased, way of examining the effects of the HGP. The present paper considers the network of collaboration between five leading participants in the HGP with Erdős numbers of 2 and six other scientists, four of whom are Nobel laureates in Physiology or Medicine, who played early, prominent and well-acknowledged scientific and leadership roles in the HGP. Collaborative links between these eleven researchers and ten other Nobel laureates in Physiology or Medicine are also described. These latter ten Nobel laureates all have Erdős numbers of at most 5; with one exception (Francis H. C. Crick); their low Erdős numbers can be traced to the HGP. As Nobel laureates are typically natural centers of subfield-specific collaborative networks, the existence of numerous Nobel laureates in Physiology or Medicine near a well-known center of collaboration in discrete mathematics provides a striking illustration of convergence between previously-distinct fields of research.

The collaborative structure of the HGP described here motivates a conjecture: that the numbers of distinct paths of each length connecting any two researchers provides a more indicative measure of the robustness of research collaboration than either minimal connecting path lengths or other single-path measures employed in analyzing collaboration graphs. Guided by this conjecture, we define in Section 4 a metric that measures all distinct paths between vertices in biconnected subgraphs and hence in the biconnected blocks of a collaboration graph. We then describe the results of employing this metric to characterize a subgraph of the Erdős Collaboration Graph containing the intuitive center of the HGP. We finally conclude that interdisciplinary collaborative links between central members of otherwise-disparate research communities may be a common feature of collaboration graphs; brokerage between disciplines or subdisciplines and hence high betweenness centrality at the scale of community structure may, in other words, be important markers of centers of collaborative effort. If this is the case, it would suggest that a multi-disciplinary education may prove increasingly valuable to both mathematicians and other scientists who wish to be near the centers of their communities.

2. An HGP subgraph of the Erdős Collaboration Graph

Figure 1 shows collaborative links, labeled by joint publications, between eleven leading participants (corresponding to bold-faced vertex labels) in the HGP, chosen on the basis of both prominence and influence over the course of the HGP. Five of these scientists have Erdős numbers of 2, conferred by publications in 1990 (MSW), 1999 (ESL), 2002 (ERK), 2003 (DJG) and 2006 (CRC) respectively; all can be considered among the founders of bioinformatics. Another four of the eleven scientists shown are Nobel laureates in Physiology or Medicine (JES, SB, JDW, HOS); however, none of these awards were made on the basis of the HGP. The other two scientists shown, JCV and FHC, were the leaders, respectively, of the two competing collaborations that simultaneously published complete "draft" sequences of the human genome in 2001. (Table 1 below provides a list of the vertex labels.)

Table 1: Vertex labels for the HGP subgraph displayed in Figure 1.

NMA	Noga M. Alon	DB	David Baltimore	CRC	Charles R. Cantor
SB	Sydney Brenner	MRC	Mario R. Capecchi	WAH	William A. Haseltine
AZF	Andrew Z. Fire	FRKC	Fan R.K. Chung	DJK	Daniel J. Kleitman
DJG	David J. Galas	FSC	Francis S. Collins	EVK	Eugene V. Koonin
ERK	Eric R. Kandel	FHCC	Francis H.C. Crick	LAM	Luc A. Montagnier
ESL	Eric S. Lander	HRH	H. Robert Horvitz	AMO	Andrew M. Odlyzko
CCM	Craig C. Mello	HOS	Hamilton O. Smith	RJR	Richard J. Roberts
JES	John E. Sulston	LAS	Laszlo A. Szekely	HEV	Harold E. Varmus
JCV	J. Craig Venter	JDW	James D. Watson	MSW	Michael S. Waterman

At most one link is shown in the graph for each pair of scientists. However, several of them have published multiple papers together (at least 50 in the case of JCV and HOS, for instance). In such cases, the earliest coauthored paper is used as the edge label. One paper reporting the draft human genome sequence [25] is the sole link between five pairs of coauthors.[2] The order-11 subgraph of the Erdős Collaboration Graph representing these HGP participants has 19 edges and contains the complete graph K_4 as a subgraph (ESL-EKV-JCV-FSC). As expected for central figures in a research community, all vertices in this subgraph have multiple additional incident edges in the full Erdős Collaboration Graph that represent joint publications with additional coauthors; some have hundreds of such incident edges.

[2]Based on searches on Google Scholar in September, 2011.

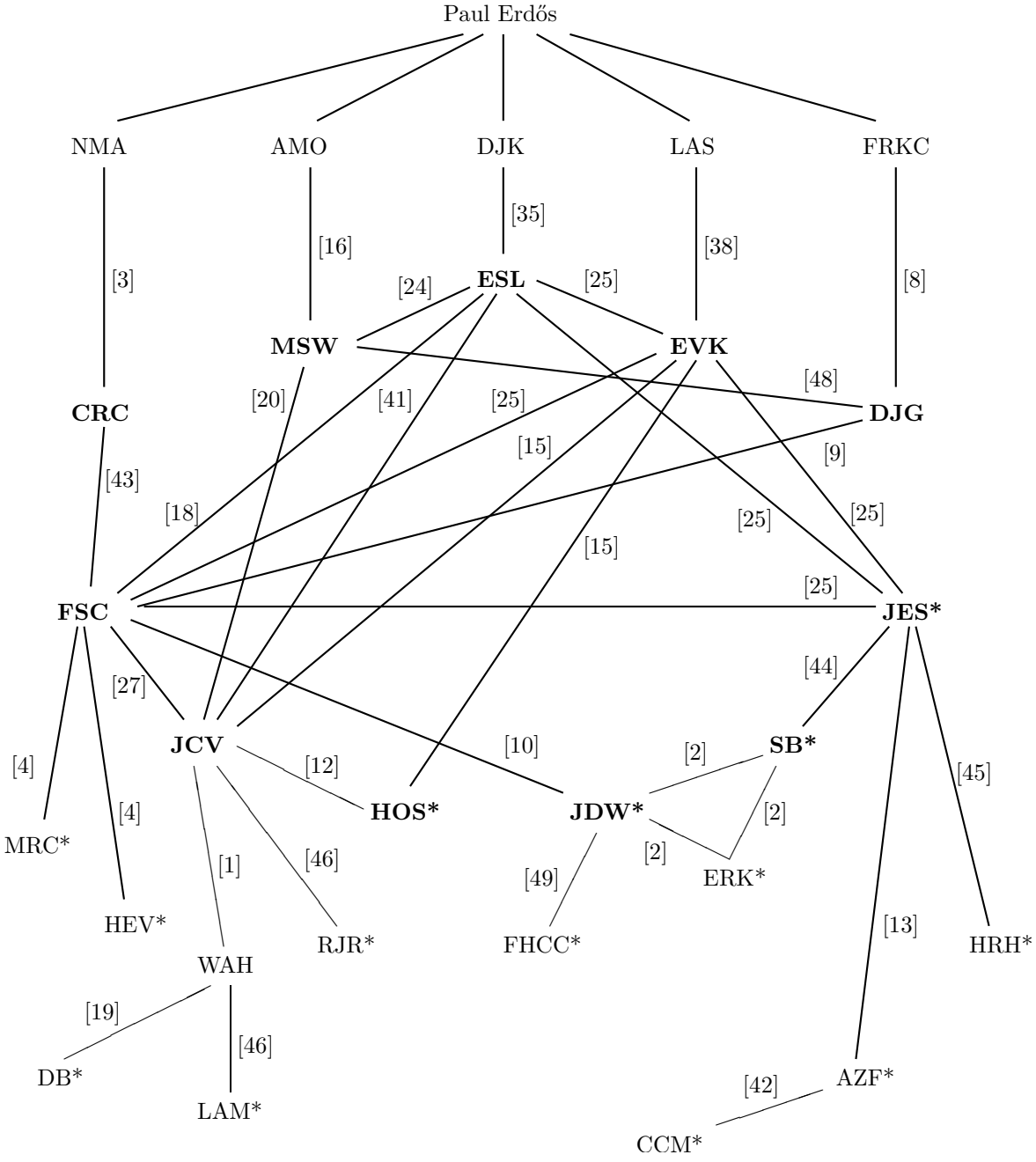

Figure 1: HGP subgraph of the Erdős Collaboration Graph. The eleven vertices with the bold-faced labels represent particularly prominent participants in the Human Genome Project; asterisks indicate Nobel laureates. Edges are labeled by selected joint publications. Vertex labels are defined in Table 1 on the previous page. Data for collaborators with Erdős are from the Erdős Number Project (http://www.oakland.edu/enp/).

The HGP' subgraph comprising the 11 bold-faced vertices shown in Figure 1 and the links between them illustrates a common feature of acknowledged centers of large-scale scientific collaboration: a kind of robustness that results from multiple collaborative links between the leading players. While one publication - [25] - links several vertices as noted, removing the edges corresponding to this publication does not disconnect the graph. The relevant scientists are, moreover, linked by multiple paths of length 2 or greater in the full Erdős Collaboration Graph; for example, JES is linked to HOS, JCV, FSC and MSW by paths of length 3 through C. Fields. The HGP subgraph contains only one bridge (FSC-CRC); however, the vertex CRC that would be disconnected by removing this bridge from the subgraph is also connected (via coauthors M. Olson, L. Hood and D. Botstein of [34]) to several other vertices of the HGP subgraph by paths of length 2 in the full Erdős Collaboration Graph.

Five vertices of the HGP subgraph are linked to Paul Erdős, and hence to an acknowledged center of discrete mathematics, by five distinct paths of length two. As noted, these paths confer Erdős numbers of 2 on five of the eleven HGP researchers shown (CRC, MSW, ESL, EVK, DJG); they confer Erdős numbers of 3 on five others (FSC, JCV, HOS, JDW, JES) and an Erdős number of 4 on SB, for an average of 2.6. This can be compared with the average, 3.3 of the Erdős numbers of Fields Medal winners from 1990 to 2010 as listed by the Erdős Number Project. Two of the papers conferring these extraordinarily low Erdős numbers ([16] and [35]) were published before the publication of the initial human genome sequence; both report bioinformatics research that contributed to the technical feasibility of the HGP. The other three ([3, 8, 38]) were published after the draft human genome sequence; all report research contributing to post-sequencing studies of gene-product function. An increasing closeness between biology and discrete mathematics as measured by Erdős numbers can, in these cases, plausibly be regarded as both enabling and enabled by the HGP.

The HGP subgraph also shows collaborative links between the eleven HGP scientists represented by the bold vertices and ten additional Nobel laureates in Physiology or Medicine. Four of these Nobel laureates (DB, ERK, CCM, and LAM) have Erdős numbers of at most five on the basis of Figure 1; the rest have Erdős numbers of at most four. The specialties of these scientists range from neuroscience (FHCC and ERK) and virology (DB and LAM) to basic cellular (MRC and HEV), developmental (AZF, HRH,

and CCM) and molecular (FHCC and RJR) biology. The HGP is, therefore, closely linked to acknowledged centers of research and collaborative activity in all of these areas. In most cases (eight out of eleven for the publications shown here as edge labels), these collaborative links were established before the 2001 publication of the initial human genome sequence.

Newman [31] reports an average distance of 4.6 between vertices of the collaboration graph representing biomedical science between 1995 and 1999; if one assumes that the average distance between authors decreases at the same rate that the average number of coauthors per paper increases, one would expect from the data of [36, 47] that the average distance between biomedical scientists in 2005 was roughly 3.5. As average distances to centers such as Nobel laureates can be expected to be smaller than average distances to more typical colleagues, one can speculate that most laboratory (as opposed to field) biologists not only have finite Erdős numbers as Grossman has suggested, but have Erdős numbers of at most eight, i.e. within the same range of most mathematicians [17]. If this is the case, it would indicate that the distance within the Erdős Collaboration Graph between the center of laboratory biology as a whole and the center of discrete mathematics represented by Erdős has decreased since the initiation of the HGP to be as close as physics has been historically. The average of the Erdős numbers of Physiology and Medicine Nobel laureates since 1970 listed by the Erdős Number Project (4.1) is, indeed, lower than the average for listed Physics Nobel laureates since 1970 (5.0); while these lists are incomplete and may reflect ascertainment bias, they also suggest that laboratory biology and physics as disciplines have similar average distances from discrete mathematics.

3. Centers and minimally-redundant blocks in collaboration graphs

The structure of the HGP subgraph shown in Figure 1, its proximity to Paul Erdős, and the close linkage of Nobel laureates in multiple subdisciplines of biology that it displays are interesting as sociological facts about the life-sciences research community and its relation to the mathematics research community. More interesting, however, is what the HGP subgraph suggests about the structure of centers of collaboration in highly-technological laboratory sciences. A center of a collaborative network can be given a traditional graph-theoretic definition as a vertex from which the greatest distance to any other vertex is the radius of the network; centers

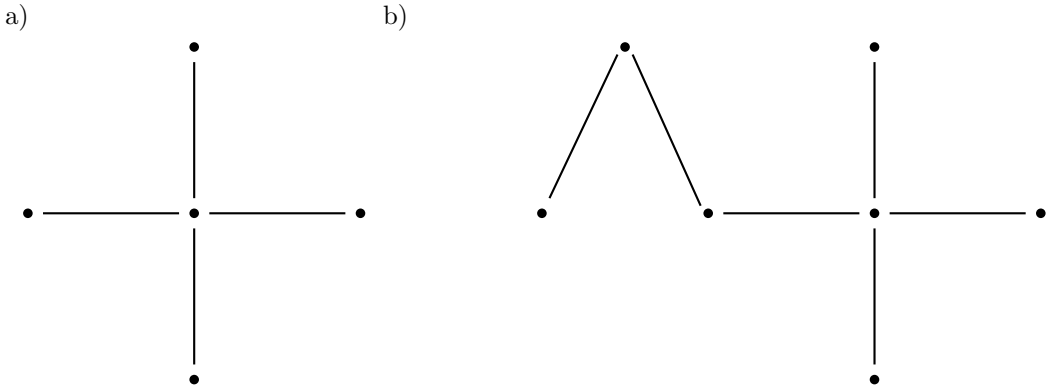

Figure 2: Relations between distance, degree, and betweenness as measures of centrality. a) All three measures coincide in a star graph; b) They do not coincide in general.

can also be characterized as vertices with maximum degree (*degree centrality*) or vertices through which maximum numbers of paths flow (*betweenness centrality*). These three measures of centrality coincide in a star graph, but as shown in Figure 2, do not coincide in general. The relevance of these measures to intuitive notions of centrality in collaboration graphs, as well as the degree to which they correspond in such graphs, is a matter of considerable debate [6, 14, 23, 26, 30, 32, 33, 39]. Consistent with previous observations of research centers, Figure 1 shows that the center of the HGP is a cluster of mutually-connected high-degree vertices. It also suggests, however, that rich connectivity to other disciplinary centers is an aspect of centrality; the HGP center is connected to Erdős, for example, not by a single path but by five distinct paths even at length two. The close relationship between the HGP, Erdős and his immediate collaborators, and Nobel laureates in other parts of biology suggests that the tradeoff between disciplinary cohesion and interdisciplinary brokerage [23] may not be a tradeoff for the highly-productive leaders of major collaborative research efforts; indeed interdisciplinary brokerage may increasingly be a requirement of community leadership.

Any connected graph, and therefore any collaboration graph, exists somewhere on the architectural spectrum [11] between a tree, in which each vertex is a cut vertex and each edge is a bridge, and a complete graph in which each vertex is connected to every other. In the case of collaboration graphs, trees represent the extreme case of communities in which a master works one-on-

one with a set of disciples, each of whom in turn works exclusively with their own sub-disciples, etc. Every member of such a community has a single, well-defined lineage: that member's single path back to the master that roots the tree. Collaborations between $(n+1)^{th}$-order disciples within the lineage of a given n^{th}-order disciple may introduce occasional cycles at distance n from the root, but do nothing to alter the essentially sectarian structure of the community as a whole. Complete graphs, on the other hand, represent communities that have no lineage structure, even locally; every member of such a community has multiple paths to every other member. Finite lifetimes, personal loyalties, and academic tradition assure that the collaboration graphs of research communities have lineage structure; practices such as finding different institutions and hence different collaborators for one's graduate training, postdoctoral work, and first faculty job tend to degrade it. It is interesting to ask, especially from the perspective of collaborations between academic disciplines, how such lineage structure degrades as the scale and scope of research collaboration increases and hence the average degree of the vertices in the collaboration graph increases toward the complete-graph limit.

The idea that *betweenness* provides a measure of centrality in collaboration graphs is motivated by the observation that some individuals hold a community together and so are traversed by multiple paths connecting individuals in one part of the community to individuals in another part [6, 26, 30]. Scientists whose collaborations bridge disciplines clearly have high betweenness centrality. Betweenness centrality alone, however, does not distinguish researchers who are *weak links*, i.e. who are the *only* connection between two or more otherwise-isolated lineages, from researchers who are central players in robust, multiply-connected collaborations like the HGP. A plausible initial step toward characterizing the robustly connected components of research communities is, therefore, to remove any lineages dependent on weak links from their collaboration graphs. This can be done by deleting bridges, thus breaking the collaboration graph into disconnected blocks. Such a move decreases the betweenness centrality of the surviving vertices, but it does not decrease the betweenness centrality within communities, and more importantly, it does not reduce the betweenness centrality that measures the significance of a vertex with respect to a robustly-collaborative multidisciplinary community. It removes from the collaboration graph papers by lone cross-disciplinary pioneers that constitute weak links, but it more clearly reveals the strong links that indicate robust interdisciplinary collaboration.

The collaboration graph of any N-author publication P, considered in isolation, is itself a complete graph K_N, with each edge labeled "P." Consider the embedding of such a graph into an existing multi-publication collaboration graph G. If at most one author of P is already represented by a vertex of G, K_N can be embedded in G with all edge labels preserved. If two or more authors of P are already represented by vertices of G, the embedding can be done in one of two uniform ways: either the edge labels already employed in G are maintained for any edges of G that are redundant with (i.e. connect the same vertices as) edges of the embedded K_N, or such labels are replaced by the label "P." In the former case only the edges in the non-redundant subgraph of K_N maintain their original "P" labels, but these labels continue to be maintained as additional publications with redundant edges are embedded in the collaboration graph. In the latter case, K_N is embedded as a labeled subgraph, but some of its edge labels may be replaced as additional publications with redundant edges are embedded in the collaboration graph. Hence with either method of managing labels on embedding, the eventual fate of any K_N representing an N-author publication P that is embedded into a collaboration graph that is dense in collaborative connections will be that some of its edges will have their "P" labels replaced by other, either older or newer, labels representing different publications.

In any collaboration graph, vertices whose edges all have a single label "P" represent authors who have collaborated only with the coauthors of some particular publication P. While the coauthors of P may have published many papers together, any author who has collaborated only with the other coauthors of P is not traversed by any shortest path joining authors of papers other than P and therefore has low betweenness centrality within the entire collaboration graph. Intuitively, such authors cannot be regarded as *central* or even *prominent* within a collaborative network regardless of their productivity, as they do not contribute to the overall connectedness of the network. Removing all such vertices from a collaboration graph does not affect its long-range connectivity. Hence one approach to identifying the vertices that are plausible candidates for centrality in the robust sense being sought here is to remove from the collaboration graph all of the vertices that correspond to authors who have only collaborated with other coauthors of a single multi-author paper. The vertices that remain are those that *do* contribute to long-range connectivity, and hence those for whom the notion of betweenness centrality is intuitively compelling.

Let G be a connected graph with unique vertex labels but possibly non-unique edge labels, and consider the subgraphs produced by the following procedure: 1) delete all bridges, breaking the graph into disconnected blocks; 2) delete all vertices from these blocks for which all the incident edges have the same label, together with their incident edges. Call a subgraph resulting from this procedure a *minimally-redundant block* in recognition of the fact that edge labels appear redundantly in such a subgraph only when the redundantly-labeled edges connect vertices required to preserve distinct edge labels. Figure 3 illustrates this procedure.

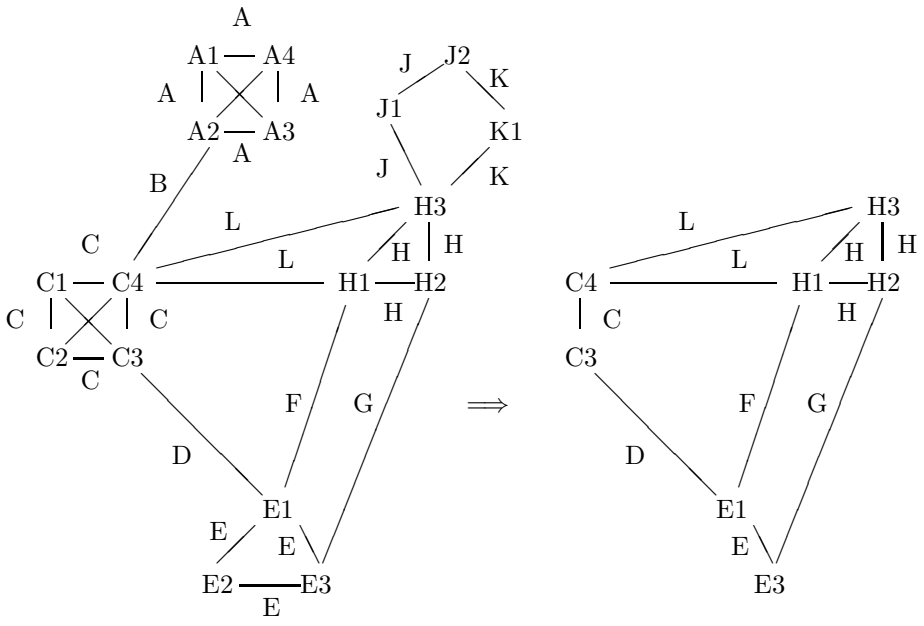

Figure 3: Construction of the minimally-redundant block of a graph. Edge labels have been added in alphabetical order, and older edge labels are kept when new edges are embedded. Labels on diagonal edges in the A and C subgraphs are suppressed for clarity.

A collaboration graph can have multiple minimally-redundant blocks of various sizes. Whether the Erdős Collaboration Graph has a vertex-number maximal minimally-redundant block is unknown; the large number of Nobel laureates with known Erdős numbers suggests that it may. It is clear, at any rate, that the Erdős Collaboration Graph does not split into minimally-redundant blocks corresponding to the traditional academic disciplines; the extent of cross-disciplinary collaboration, even in the 1970s, was too large for this to occur.

In the context of collaboration graphs, the identification of a minimally-redundant block has two main effects. First, it removes from the collaboration graph all lineages dependent on weak links. Second, it reduces the degree of most if not all high-degree vertices, and altogether removes high-degree vertices representing authors who have collaborated only with other coauthors of a single multiple-author paper, regardless of their total productivity. As an example, the largest minimally-redundant block of the subgraph shown in Figure 1 consists of all vertices and edges other than those below the bold-faced vertices: the deletion of bridges in step 1 of the procedure removes all vertices below the bold-faced ones except ERK, which is removed in step 2 because its two incident edges have the same label. The vertices deleted in this construction would all be preserved in a minimally-redundant block constructed from the complete Erdős Collaboration Graph due to the existence of additional paths; there are, for example, paths of length three connecting each of MSW, FSC, JCV and HOS to either AZF or CCM through C. Fields. On the other hand, a minimally-redundant block containing Erdős constructed from the full Erdős Collaboration Graph would not include the vertices representing a number of individuals with Erdős numbers of one but no coauthors other than Erdős (e.g. N. Anning or D. T. Busolini, given the data of the Erdős Number Project). In the region of the graph representing the HGP, constructing such a minimally-redundant block would delete a number of vertices with degree greater than 100 but no coauthors beyond those of a single paper (e.g. S. Wenning or D. Hostin, coauthors of [25] and [46], respectively[3]) and significantly reduce the degrees of many other high-degree vertices. Such effects would also be expected in other areas involving big science collaborations, such as experimental high-energy physics.

4. A conductance metric for minimally-redundant blocks

Consider two distinct vertices x and y of a graph G with unique vertex labels and possibly non-unique edge labels. List all paths from x to y, ordered by their length i from the smallest value of i to the largest, with an arbitrary order for paths of equal length. Delete from this list, in order from the smallest value of i, all paths containing two or more edges with the same label, and all paths containing any edge with the same label as any edge

[3]Based on searches on Google Scholar on 11 October 2011.

contained in any path appearing earlier in the list. Then define a function $\zeta(x, y) = (\sum_i (p_i/i))^{-1}$, where (1) terms are added in order of increasing i, and (2) p_i is the number of paths from x to y of length i that remain on the list, i.e. paths with distinct edge labels that share no edge labels with any paths already counted.

If x is a leaf vertex of a tree that is separated from the root y by n edges, all with distinct labels, $\zeta(x, y) = n$; if the path from x to y contains duplicate labels, it cannot be counted and $\zeta(x, y)$ is undefined. If x is a vertex in K_N, then for any other vertex y, $\zeta(x, y) = 1$ if the edges all have the same label, as only the single path of length 1 from x to y can be counted without re-using edge labels. If x and y are distinct vertices in K_N and all edge labels are distinct, the path of length 1 from x to y can be counted, as can the paths of length 2 from x to y via each of the remaining $N-2$ vertices. All paths of greater length re-use edge labels and cannot be counted; hence in this case $\zeta(x, y) = (1 + (N-2)/2)^{-1} = 2/N$.

The construction of $\zeta(x, y)$ nonlinearly rewards pairs of vertices x and y that are connected by multiple paths, even if x and y are adjacent; however, it does so only for paths that do not introduce edge-label redundancy. In the context of a coauthorship graph, $\zeta(x, y)$ nonlinearly rewards pairs of authors who have colleagues who are connected at any degree of collaboration (e.g. by collaborations between colleagues of colleagues), provided that these collaborations produced not-yet-counted publications. For example, in the Erdős Collaboration Graph, the values of $\zeta(x, y)$ where y represents Erdős range from less than 0.04 if x is F. Harary to values on the order of 20, assuming that at least some non-mathematicians have finite Erdős numbers larger than those of any mathematicians. The dynamic range of $\zeta(x, y)$ for the Erdős Collaboration Graph is therefore on the order of 500, an improvement in dynamic range of around 25 times over minimal path length. Similar improvements in dynamic range would be expected for any large social network in which the degree distribution approximately follows a power law. This increased dynamic range would, if $\zeta(x, y)$ could be employed as a metric, result in a much more sensitive measure of closeness between x and y than that provided by measures such as minimal path length.

In the case of arbitrary graphs with distinct edge and vertex labels, however, the function $\zeta(x, y)$ fails to satisfy the triangle inequality and is therefore not a metric. For example, in the graph shown in Figure 4, $\zeta(a, b) = 1$,

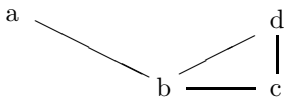

Figure 4: An example showing that $\zeta(x, y)$ is not a metric for arbitrary graphs.

$\zeta(b, c) = 2/3$, and $\zeta(a, c) = 2 > 1\ 2/3$ because the $a - b$ edge can only be counted once. However, for biconnected graphs, $\zeta(x, y)$ is a metric. To see this, note that $\zeta(x, y)$ is a metric for complete graphs, with K_3 as the simplest case. For induction, consider a graph comprising a biconnected subgraph M for which $\zeta(x, y)$ is a metric together with one additional vertex a that is adjacent to two distinct vertices b and c of M. In this case, the values $\zeta(a, b)$ and $\zeta(a, c)$ must be equal, so $\zeta(a, c) \leq \zeta(a, b) + \zeta(b, c)$ regardless of the value of $\zeta(b, c)$. It is clear, moreover, that an additional edge can be added to any biconnected graph without violating the triangle inequality, as adding an edge will decrease the value of $\zeta(x, y)$ for the two vertices it joins more than for any other two vertices.

The biconnected graph metric $\zeta(x, y)$ is a special case of Kirchhoff's well-known metric for conductance in parallel electrical circuits. Because it combines information about path lengths, vertex degrees, and betweenness, employing it as a metric in collaboration graphs reveals structure that is evident in high-resolution illustrations but difficult to see at low resolution or with other metrics. In the subgraph shown in Figure 1, for example, ζ(ESL,Erdős) $\simeq$ 0.71 while ζ(CRC,Erdős) $\simeq$ 1.33 even though ESL and CRC both have Erdős numbers of 2. Similarly, ζ(JCV,Erdős) = 0.80 while ζ(HOS,Erdős) $\simeq$ 1.88 even though JCV and HOS both have Erdős numbers of 3. The differentiation obtained with $\zeta(x, y)$ is even greater in other parts of the Erdős Collaboration Graph. For example, considering only paths up to length 2, $\zeta(x$,Erdős) $\simeq$ 0.11 for $x =$ L. Lesniak, but $\zeta(x$,Erdős) = 1.0 for $x =$ E. Lesigne, even though both authors have Erdős numbers of 2. Considering only paths up to length four, $\zeta(x$,Erdős) = 0.46 for $x =$ C. Fields, with Erdős number 3.[4]

[4]Values for C. Fields were computed using data available at http://chrisfieldsresearch.com/erdos.htm, accessed on June 24, 2014.

5. Conclusion

As documented in [17], both collaboration in mathematics and the fraction of collaborative mathematical papers with more than two authors have been steadily increasing since the 1940s. These trends render the Erdős Collaboration Graph both increasingly interesting and, to the extent that papers with more than two authors contribute edges with redundant labels, increasingly messy with each passing decade. Extending the Erdős Collaboration Graph to include collaborations outside of mathematics makes it even more interesting, and particularly in areas that involve big science collaborations, enormously messier. One response to this increasing messiness is to restrict the graph to two-author papers, and hence to authors with Erdős numbers of the second kind. While they still outnumber one-author papers [36], however, two-author papers are relatively rare in the sciences; recent issues of *Science, Nature, Proceedings of the National Academy of Sciences of the USA* and *Physical Review Letters*, for example, have 16 two-author papers out of 175 total papers (9.1%) between them.[5] Restricting the collaboration graph to two-author papers thus enormously under-represents collaboration; in particular, it under-represents the impact of mathematics on science. It also fails to capture one of the most salient features of collaborative networks in the sciences: the robustness that results from the tendency of prominent scientists to collaborate with each other both within and between disciplinary communities.

This paper suggests an alternative way of dealing with the messiness introduced by redundant edge labels, one that not only preserves but highlights the robustness of collaborative connections both within and between research communities. Constructing the minimally-redundant blocks of a collaboration graph trims away both vertices and edges that do not contribute to a robustly-connected structure. Employing $\zeta(x, y)$ as a metric on the constructed minimally-redundant blocks reveals pairs of authors who are connected by large numbers of non-redundant paths in the collaboration graph. Using these techniques, it becomes clear that despite their differences in interests and methods from pure mathematics, contemporary molecular, cellular, and developmental biology can plausibly be considered to be math-

[5] The issues canvassed are *Science* 345(6068), *Nature* 482(7383), *Proc. Natl. Acad. Sci. USA* 109(5) and *Phys. Rev. Lett.* 108(5).

ematical sciences to the same extent that physics is. It seems likely from the Erdős Number Project's survey of Erdős numbers of prominent scientists that many of the other special sciences can be viewed in the same way. If this is correct, it would suggest that a greater emphasis on interdisciplinary knowledge and communication skills, as well as a greater exposure to mathematics, may be beneficial in graduate programs across the sciences.

References

[1] M. D. Adams, A. R. Kerlavage, R. D. Fleischman. J. Bult, N. H. Lee, E. W. Kirkness, K. G. Weinstock, J. D. Gocayne, O. White, G. Sutton and others, "Initial assessent of human gene diversty and expression patterns based upon 83 million nucleotides of cDNA sequence," *Nature*, Volume **377** Number 6547 Supplement (1995), pages 3-174.

[2] H. Akil, S. Brenner, E. Kandel, K. S. Kendler, M.-C. King, E. Scolnick, J. D. Watson and H. Y. Zoghbi, "The future of psychiatric research: Genomes and neural circuits," *Science*, Volume **327** Number 5973 (2010), pages 1580-1581.

[3] N. Alon, V. Asodi, C. Cantor, S. Kasif and J. Rachlin, "Multi-node graphs: A framework for multiplexed biological assays," *Journal of Computational Biology*, Volume **13** Number 10 (2006), pages 1659-1672.

[4] C. P. Austin, J. F. Battey, A. Bradley, M. Bucan, M. Capecchi, F. S. Collins, W. F. Dove, G. Duyk, S. Dymecki, J. T. Eppig and others, "The knockout mouse project," *Nature Genetics*, Volume **36** Number 9 (2004), pages 921-924.

[5] A.-L. Barabási, "Network theory - The emergence of the creative enterprise," *Science*, Volume **308** Number 5722 (2005), pages 639-641.

[6] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social Networks*, Volume **28** Number 4 (2006), pages 466-484.

[7] R. S. Burt, *Brokerage and Closure: An Introduction to Social Capital*, Oxford University Press, New York, 2005.

[8] F. Chung, L. Lu, T. G. Dewey and D. J. Galas, "Duplication models for biological networks," *Journal of Computational Biology*, Volume **10** Number 5 (2003), pages 677-687.

[9] F. S. Collins and D. J. Galas, "A new five-year plan for the U.S. Human Genome Project," *Science*, Volume **262** Number 5130 (1993), pages 43-46.

[10] F. S. Collins and J. D. Watson, "Genetic discrimination: Time to act," *Science*, Volume **302** Number 5646 (2003), page 745.

[11] R. Diestel, *Graph Theory* (4th Ed.), Springer, Berlin, 2010.

[12] R. D. Fleischmann, M. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and others, "Whole genome random shotgun sequencing and assembly of *Haemophilus influenzae* Rd," *Science*, Volume **269** Number 5223 (2005), pages 496-508.

[13] J. T. Fleming, M. D. Squire, T. M. Barnes, C. Tornoe, K. Matsuda, J. Ahnn, A. Fire, J. E. Sulston, E. A. Barnard, D. B. Satelle and J. A. Lewis, "*Caenorhabditis elegans* levamisole resistance genes *lev-1, unc-29* and *unc-38* encode functional nicotinic acetylcholine receptor subunits," *Journal of Neuroscience*, Volume **17** Number 15 (1997), pages 5843-5857.

[14] L. C. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks*, Volume **1** Number 3 (1978/79), pages 215-239.

[15] M. J. Gardner, H. Tettelin, D. J. Carucci, L. M. Cummings, L. Aravind, E. V. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii and others, "Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*," *Science*, Volume **282** Number 5391 (1998), pages 1126-1132.

[16] J. R. Griggs, P. Hanlon, A. M. Odlyzko and M. S. Waterman, "On the number of alignments of k sequences," *Graphs and Combinatorics*, Volume **6** Number 2 (1990), pages 133-146.

[17] J. W. Grossman, "Patterns of research in mathematics," *Notices of the AMS*, Volume **52** Number 1 (2005), pages 35-41.

[18] J. G. Hacia, J.-B. Fan, O. Ryder, J. Lin, K. Edgemon, G. Ghandour, R. A. Mayer, B. Sun, L. Hsie, C. M. Robbins and others, "Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays," *Nature Genetics*, Volume **22** Number 2 (1999), pages 164-167.

[19] W. A. Haseltine, D. G. Kleid, A. Panet, E. Rothenberg and D. Baltimore, "Ordered transcription of RNA tumor virus genomes," *Journal of Molecular Biology*, Volume **106** Number 1 (1976), pages 109-131.

[20] S. Istrail, G. G. Sutton, L. Florea, A. L. Halpern, C. M. Mobarry, R. Lippert, B. Walenz, H. Shatkay, I. Dew, J. R. Miller and others, "Whole-genome shotgun assembly and comparison of human genome assemblies," *Proceedings of the National Academy of Sciences USA*, Volume **101** Number 7 (2004), pages 1916-1921.

[21] B. Karrer, E. Levina and M. E. J. Newman, "Robustness of community structure in networks," *Physical Review E*, Volume **77** Number 4 (2008), article # 046119.

[22] R. Klavans and K. W. Boyack, "Toward a consensus map of science." *Journal of the American Society for Information Science and Technology*, Volume **60** Number 3 (2009), pages 455-476.

[23] R. Lambiotte and P. Panzarasa, "Communities, knowledge creation and information diffusion," *Journal of Informetrics*, Volume **3** Number 3 (2009), pages 180-190.

[24] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: A mathematical analysis," *Genomics*, Volume **2** Number 3 (1988), pages 231-239.

[25] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh and others, "Initial sequencing and analysis of the human genome," *Nature*, Volume **409** Number 6822 (2001), pages 860-921.

[26] A. Landherr, B. Friedl and J. Heidemann, "A critical review of centrality measures in complex networks," *Business Information Systems Engineering*, Volume **2** Number 6 (2010), pages 371-385.

[27] W. R. McCombie, A. Martin-Gallardo, J. D. Gocayne, M. FitzGerald, M. Dubnick, J. M. Kelley, L. Castilla, L. I. Liu, S. Wallace, S. Trapp and others, "Expressed genes, *Alu* repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3," *Nature Genetics*, Volume **1** Number 5 (1992), pages 348-353.

[28] F. Moya-Anegón, B. Vargas-Quesada, Z. Chinchilla-Rodríguez, E. Correra-Álvarez, F. J. Munoz-Fernández and V. Herrero-Solano, "Visualizing the marrow of science," *Journal of the American Society for Information Science and Technology*, Volume **58** Number 14 (2007), pages 2167-2179.

[29] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences USA*, Volume **98** Number 2 (2001), pages 404-409.

[30] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, Volume **45** Number 2 (2003), pages 167-256.

[31] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences USA*, Volume **101** Supplement 1 (2004), pages 5200-5205.

[32] M. E. J. Newman, "Who is the best connected scientist? A study of scientific coauthorship networks," in *Complex Networks*, E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai (eds.), Springer, Berlin, 2004, pages 337-370.

[33] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences USA*, Volume **103** Number 23 (2006), pages 8577-8582.

[34] M. Olson, L. Hood, C. Cantor and D. Botstein, "A common language for physical mapping of the human genome," *Science*, Volume **245** Number 4925 (1989), pages 1434-1435.

[35] L. Pachter, S. Batzoglou, V. I. Spitkovsky, E. Banks, E. S. Lander, D. J. Kleitman and B. Berger, "A dictionary-based approach for gene annotation," *Journal of Computational Biology*, Volume **6** Number 3-4 (1999), pages 419-430.

[36] A. L. Porter and I. Rafols, "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time," *Scientometrics*, Volume **81** Number 3 (2009), pages 719-745.

[37] L. Roberts, R. J. Davenport, E. Pennisi and E. Marshall, "A history of the Human Genome Project," *Science*, Volume **291** Number 5507 (2001), page 1195.

[38] I. B. Rogozin, K. S. Makarova, J. Murvai, E. Czabarka, Y. I. Wolf, R. L. Tatusov, L. A. Szekely and E. V. Koonin, "Connected gene neighborhoods in prokaryotic genomes," *Nucleic Acids Research*, Volume **30** Number 10 (2002), pages 2212-2223.

[39] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences USA*, Volume **105** Number 4 (2008), pages 1118-1123.

[40] I. Samoylenko, T.-C. Chao, W.-C. Liu and C.-M. Chen, "Visualizing the scientific world and its evolution," *Journal of the American Society for Information Science and Technology*, Volume **57** Number 11 (2006), pages 1461-1469.

[41] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. A. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek and others, "A gene map of the human genome," *Science*, Volume **274** Number 5287 (1996), pages 540-546.

[42] G. Seydoux, C. C. Mello, J. Pettitt, W. B. Wood, J. R. Priess and A. Fire, "Repression of gene expression in the embryonic germ lineage of *C. elegans*," *Nature*, Volume **382** Number 6593 (1996), pages 713-716.

[43] C. M. Smith, S. K. Lawrance, G. A. Gillespie, C. R. Cantor, S. M. Weissman and F. S. Collins, "Strategies for mapping and cloning macroregions of mammalian genomes," *Methods in Enzymology*, Volume **151** (1987), pages 461-489.

[44] J. E. Sulston and S. Brenner, "The DNA of *Caenorhabditis elegans*," *Genetics*, Volume **77** Number 1 (1974), pages 95-104.

[45] J. E. Sulston and H. R. Horvitz, "Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*," *Developmental Biology*, Volume **56** Number 1 (1977), pages 110-156.

[46] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt and others, "The sequence of the human genome," *Science*, Volume **291** Number 5507 (2001), pages 1304-1351.

[47] M. L. Wallace, V. Larivière and Y. Gingras, "A small world of citations? The influence of collaboration networks on citation practices," *PLoS One*, Volume **7** Number 3 (2012), article # e33339.

[48] M. S. Waterman, R. Arratia and D. J. Galas, "Pattern recognition in several sequences: Consensus and alignment," *Bulletin of Mathematical Biology*, Volume **48** Number 4 (1984), pages 515-527.

[49] J. D. Watson and F. R. C. Crick, "Molecular structure of nucleic acids," *Nature*, Volume **171** Number 4356 (1953), pages 737-738.

[50] J. D. Watson and R. M. Cook-Deegan, "Origins of the Human Genome Project," *FASEB Journal*, Volume **5** Number 1 (1991), pages 8-11.