An Empirical Study of CSSS 2013 Participant-Project Network with Participants' Academic Background as "Genotypes"

Yan Xu （许 晏）

Department of Physics, Florida State University, Tallahassee, United States

Email: yxu2@fsu.edu or ming.hsuyen@gmail.com

Puduru Viswanadha Reddy

Groupe d'études et de recherche en analyse des décisions (GERAD),

École des Hautes Études commerciales de Montréal (HEC Montréal), Montréal, Canada

Email: puduru.reddy@gerad.ca or viswanadha.puduru@gmail.com

September 16, 2013

1. **Introduction**

In this brief report, we investigate the structure of participant-project network formed during the 2013 Complex Systems Summer School (CSSS) at Santa Fe. The data is based on [1], telling us who works with whom in which project. We have also collected data of participants' academic background [2]. Each participant's background is expressed by a 4-component vector ("gene"), representing 4 different categories of subjects in the following order: math & physics, life sciences & ecology, social sciences & economics, and computing & programming. For each category, we ask a participant the following question: "Have you ever taken any graduate level courses or had research experience in that subject?" If the answer is "Yes", we assign "1" to the component corresponding to that subject; otherwise, "0" is assigned to that component. For example, a participant with "genotype" (1 0 1 0) means that he has graduate level background in both math & physics and social sciences & economics, but none in life sciences & ecology nor computing & programming. In this way, a 4-component genotype vector characterizing academic or research background is associated with each participant. Since the nature of this summer school is transdisciplinary, it is of intrinsic interest to study how various "genotypes" are mixed in the CSSS collaboration network, which may show interesting pattern or complexity.

2. The CSSS 2013 Participant-Project Bipartite Network

We represent the 61 participants and 20 projects in [1] as a bipartite graph (see Figure 1), in which vertices are divided into two disjoint sets X (participants) and Y (projects) such that every edge connects a vertex in X to one in Y. When a participant vertex $x \in X$ is connected to a project vertex $y \in Y$, it simply means that participant x works on project y. Note that multiple participants may work together in the same project, and one participant may also work on multiple projects. In fact, out of the total 61 participants, the numbers of participants who work on one, two and three projects are 34, 20 and 7, respectively.

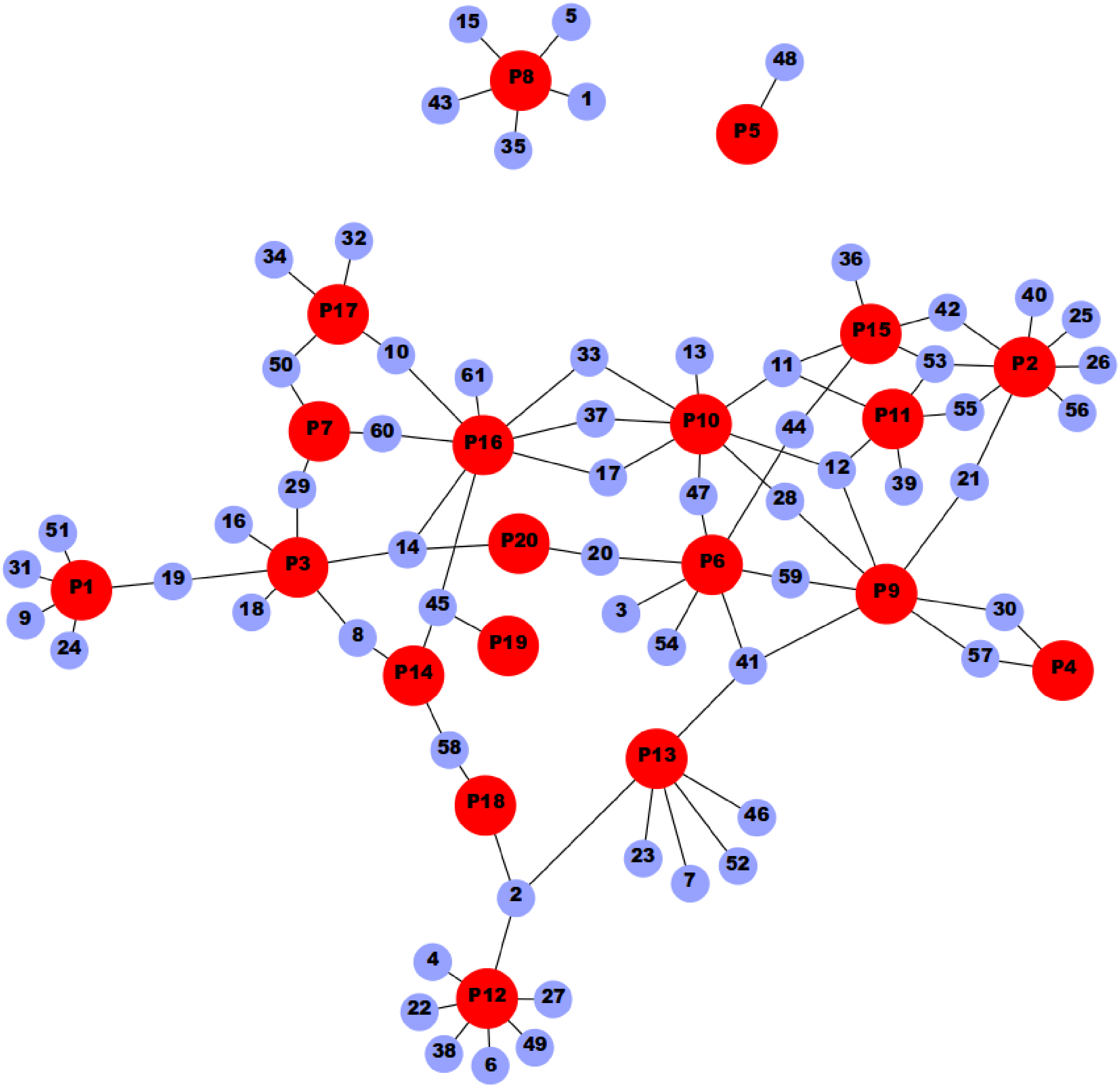

Figure 1. The bipartite network representation of CSSS 2013 participants (blue vertices) and projects (red vertices). Every edge connects a participant with a project.

3. The CSSS Network of Projects

A bipartite network is also called a two-mode network in the sociology literature (see [3], pp. 123-126). We can use the bipartite network to infer connections *within* participants or projects, by creating a one-mode projection from the two-mode bipartite form. For example, when it is projected onto the vertices of "projects" (or "groups"), we obtain a one-mode network of projects (see Figure 2a), in which vertices are CSSS projects, and two projects are connected if there is at least one participant working on both during the summer school. In Figure 2b, we show the distribution of group size in histogram. The average group size is 4.75, i.e., there are in average 4.75 participants in each project group. We also analyze the vertex degree distribution for the network of projects in Figure 3a. Notice in Figure 2a, two projects (P5 and P8) are isolated from the giant component of projects. Groups of larger sizes usually have higher degrees, e.g., P16.

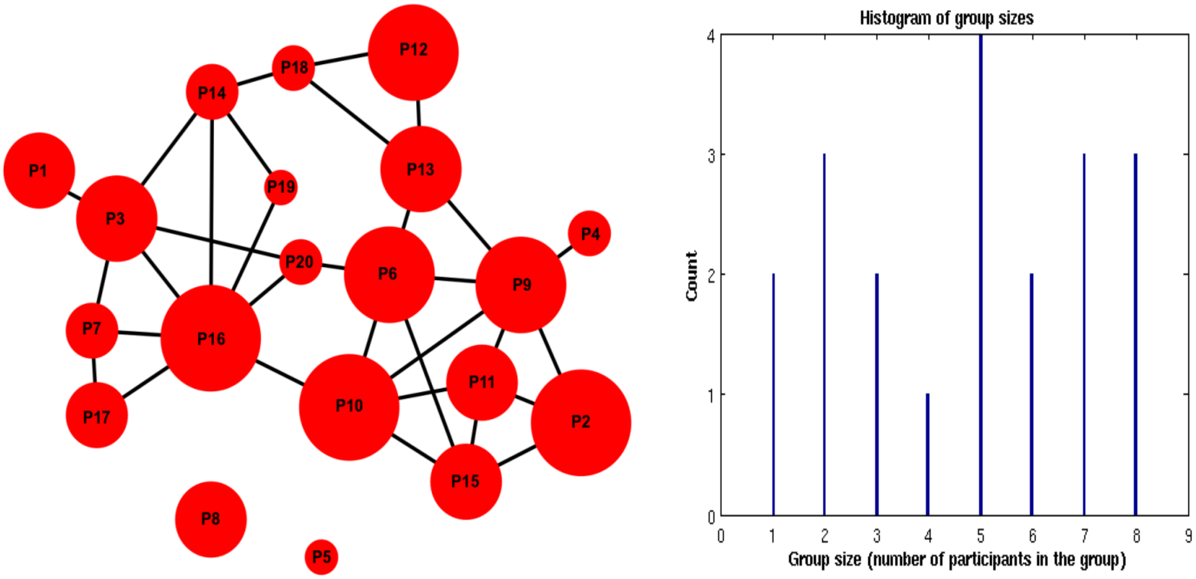

Figure 2a

Figure 2b

Figure 2. (a) The one-mode projection of CSSS 2013 bipartite network onto projects. Two vertices (projects) are connected by an edge if there are participants working on both of them. The size of each vertex is proportional to the group size, i.e., number of participants in the project group. (b) Distribution of project group size in histogram.

4. The CSSS collaboration network of participants with "genotypes"

In this section, we study the one-mode projection of the participant-project bipartite network onto participants, which is our main interest of study in this report. In this CSSS collaboration network (see Figure 4), vertices are participants, and two participants are connected by an edge if they collaborate within a project. Figure 3b shows the degree distribution of this network.

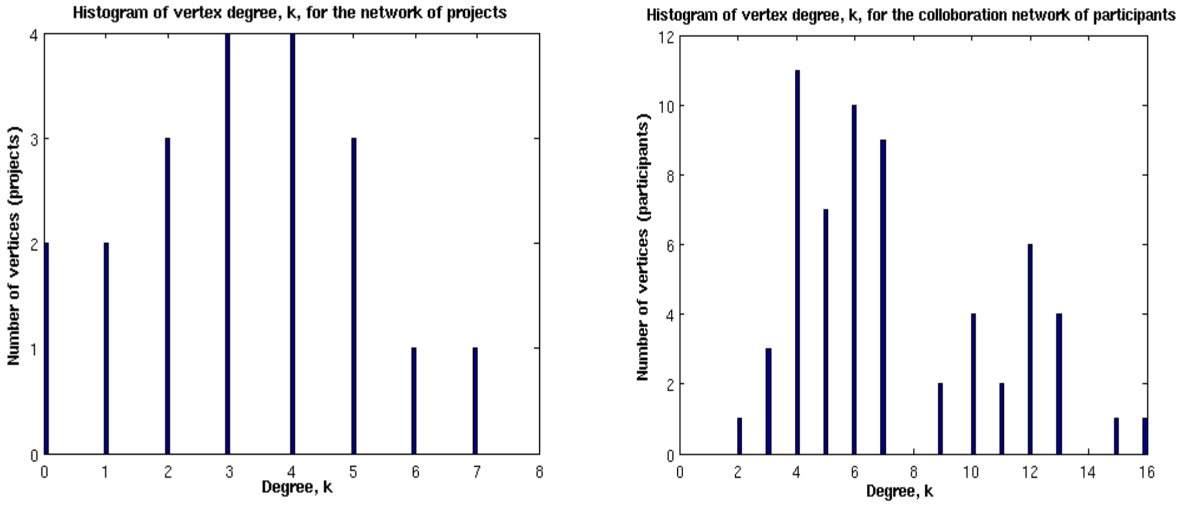

Figure 3a

Figure 3b

Figure 3. (a) Vertex degree distribution for the network of projects. (b) Vertex degree distribution for the network of participants.

We can see clique subgraphs (indicating community structure) in Figure 4. This is due to the fact that within the same project group, every participant is connected to every other member of that group, thus forming a complete network as a subgraph in the whole participant network. Recall that in Figure 2a, two projects (P5 and P8) are isolated from the giant component in the project network, hence we see in Figure 4 these two groups of participants are also isolated from the giant component in the participant network. The whole participant network has 61 vertices and 224 edges, so the average degree is about 7.34. In the giant component, there are 55 vertices and 213 edges, thus the average vertex degree is about 7.75; the diameter (largest vertex distance) is 6 and the average path length (mean vertex distance) is around 2.88.

Each vertex (participant) is associated with a 4-component "genotype" vector representing academic or research background. Define P_α to be the fraction of vertices with genotype α, and the probability distribution $\{P_\alpha\}$ of all possible genotypes $\{\vec{V}_\alpha\}$ in CSSS participants is

$\vec{V}_\alpha$	$\begin{pmatrix}1\\0\\0\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\\1\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\\1\end{pmatrix}$	$\begin{pmatrix}0\\1\\1\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\\0\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\\1\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\1\\1\end{pmatrix}$	$\begin{pmatrix}0\\1\\1\\1\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\\0\\0\end{pmatrix}$
P_α	$\frac{20}{61}$	$\frac{8}{61}$	$\frac{7}{61}$	$\frac{7}{61}$	$\frac{4}{61}$	$\frac{3}{61}$	$\frac{3}{61}$	$\frac{2}{61}$	$\frac{2}{61}$	$\frac{2}{61}$	$\frac{1}{61}$	$\frac{1}{61}$	$\frac{1}{61}$

The Shannon entropy of such probability distribution is $S = -\sum_\alpha P_\alpha \log_2 P_\alpha \approx 3.09$.

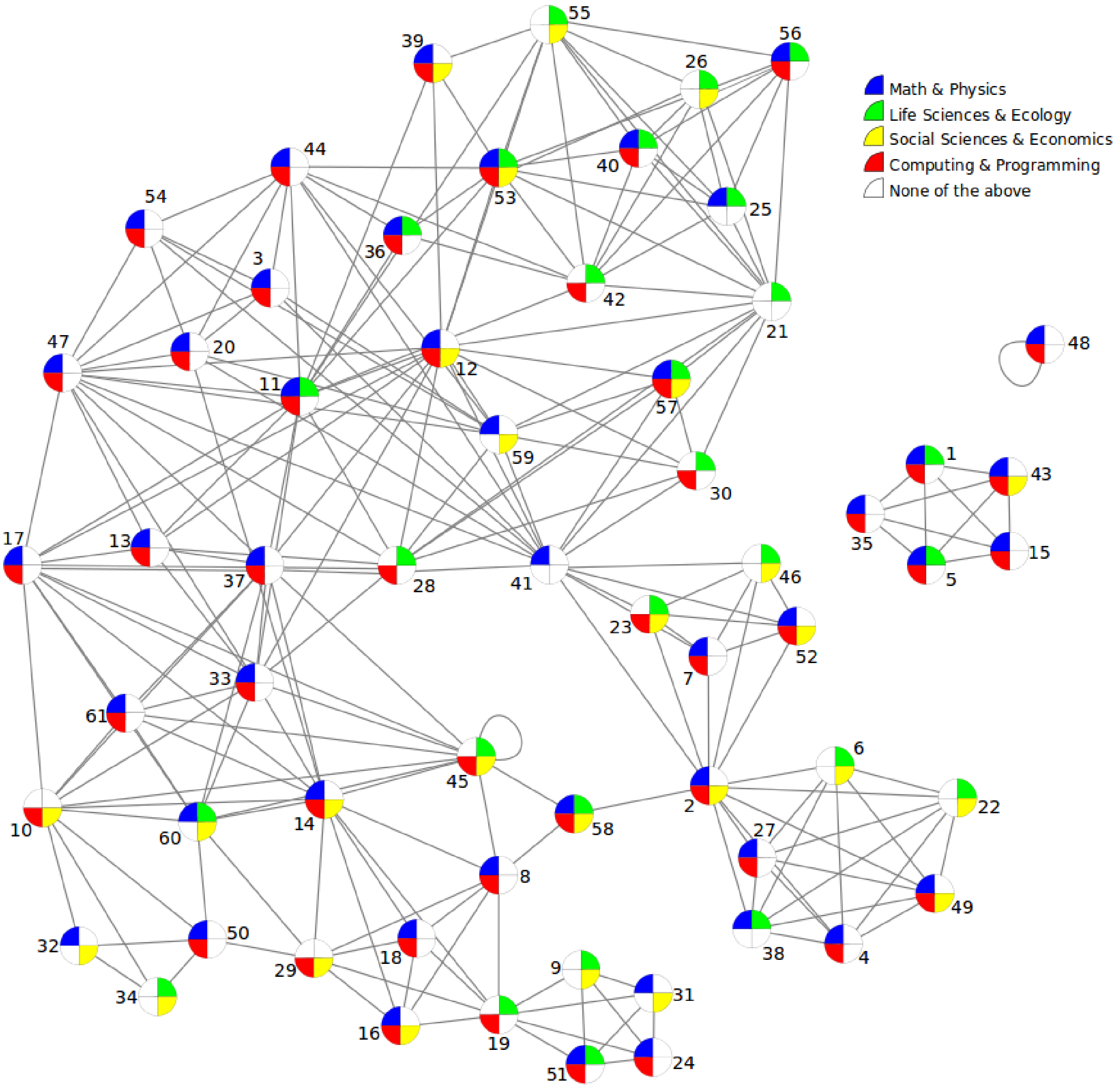

Figure 4. The one-mode projection of CSSS 2013 bipartite network onto participants. Two vertices (participants) are connected by an edge if they have project collaboration with each other. Notice there are two self-edges in the network, one for vertex No. 45 and one for vertex No. 48, indicating that each involves in a single-person project, which can also be seen in the bipartite network representation. Each vertex i is associated with a 4-component "genotype" vector $\vec{V}_i$ representing academic or research background. We use four different colors (blue, green, yellow and red) to denote the four different categories of subjects, as illustrated in the inset of this figure.

5. Betweenness centrality analysis in the giant component of the network of participants

In the giant component of the network of participants, besides communication or discussion between participants inside each project group, information also flows and is shared between nonadjacent vertices via paths in networks. If we assume that every pair of vertices exchanges information and that messages always take the shortest (geodesic) path through the network, betweenness centrality is a measure of the extent to which a vertex lies on paths between *other* vertices. Thus betweenness can be viewed as an approximate guide to the influence vertices (participants) have over the flow of information between others. Here we follow the definition in [3], viz. the betweenness centrality of vertex i is $b_i \equiv \sum_{st} \frac{n_{st}^i}{g_{st}}$, where n_{st}^i is the number of shortest paths from vertex s to t that pass through i, and g_{st} is the total number of shortest paths from s to t. We measure the betweenness for each vertex in the participant network, and plot the result as a function of degree in Figure 5a. Note that 30 vertices in the giant component have zero betweenness, which implies that about 55% of the participants in the giant component do not play a significant role of message passing between participants other than themselves (of course, such result is conditioned on the assumption that information flow only takes shortest paths in the collaboration network, which may not be accurate in the real world). This is due to the fact that 28 participants in the giant component work on only a single project. For a vertex to have high betweenness, high degree is not a necessary condition [4]. We analyze the distribution of betweenness normalized by vertex degree, and see some "anomalous" vertices with low degree but of relatively high betweenness. The origin of such a behavior is from participants who work on multiple projects, sitting on the overlap position of multi-communities in the network.

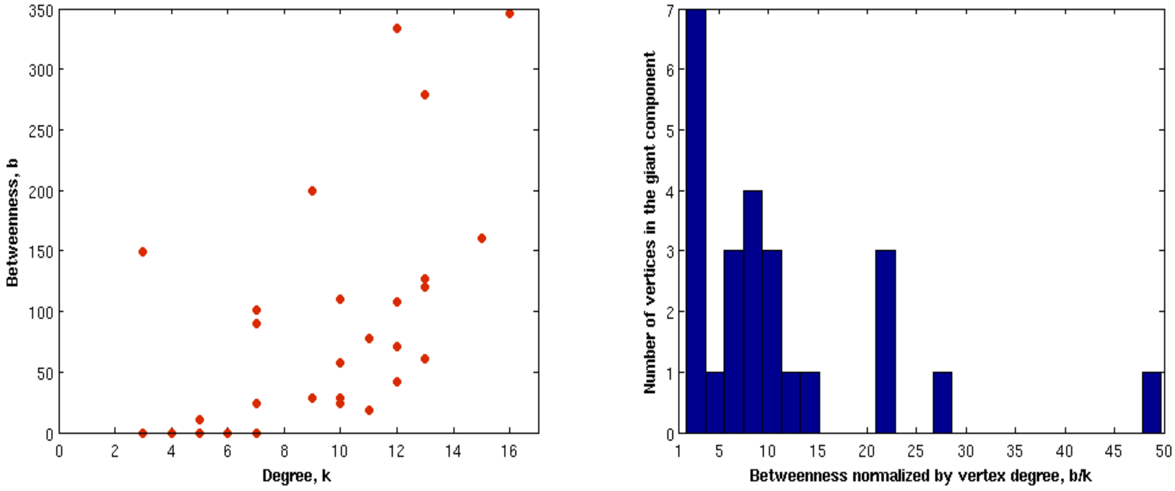

Figure 5a

Figure 5b

Figure 5. (a) Betweenness as a function of degree in the giant component of the collaboration network of participants. (b) Distribution of betweenness normalized by vertex degree. Notice the average normalized betweenness is about 5.1 and the highest value is 49.7 (vertex No. 58).

6. Cosine similarity of genotypes in the CSSS collaboration network of participants

It is interesting to investigate how participants with diverse academic or research background are structured into project groups. We introduce a cosine similarity measure between one genotype α and another genotype β. Recall that a participant's genotype is characterized as a 4-component vector, thus the genotype cosine similarity is defined as

$$\cos\theta_{\alpha\beta} = \frac{\vec{V}_\alpha \cdot \vec{V}_\beta}{\|\vec{V}_\alpha\|\|\vec{V}_\beta\|}$$

where $\|\vec{V}_\alpha\| \equiv \sqrt{V_{\alpha,1}^2 + V_{\alpha,2}^2 + V_{\alpha,3}^2 + V_{\alpha,4}^2}$ is the Euclidean norm of vector $\vec{V}_\alpha$ in the 4D genotype space, and $\vec{V}_\alpha \cdot \vec{V}_\beta \equiv V_{\alpha,1}V_{\beta,1} + V_{\alpha,2}V_{\beta,2} + V_{\alpha,3}V_{\beta,3} + V_{\alpha,4}V_{\beta,4}$ is the inner product of $\vec{V}_\alpha$ and $\vec{V}_\beta$. Since no component in the genotype vector is negative, $\cos\theta_{\alpha\beta} \in [0,1]$. Larger $\cos\theta_{\alpha\beta}$ value indicates more similarity between genotype α and genotype β.

For each edge in the network of participants (Figure 4), there is a pair of genotypes associated with the vertices at the two ends of that edge, thus there is a cosine similarity (correlation) measure associated with each edge. We measure the fraction $P_{\alpha\beta}$ of edges that connect a vertex of genotype α with a vertex of genotype β in the collaboration network. The average cosine similarity (genotype correlation) for the whole network is calculated as

$$\langle\cos\theta\rangle_{\text{real}} = \sum_\alpha \sum_\beta P_{\alpha\beta}\cos\theta_{\alpha\beta} \approx 0.644$$

For comparison, we also estimate the average cosine similarity if the population of participants with the same distribution of genotypes are randomly mixed. Recall P_α is defined to be the fraction of genotype α in the population of all 61 participants, then in a randomly mixed network, the probability of an edge joining a vertex of genotype α with a vertex of genotype β is just the product of the two marginal probabilities P_α and P_β, hence

$$\langle\cos\theta\rangle_{\text{random}} = \sum_\alpha \sum_\beta P_\alpha P_\beta\cos\theta_{\alpha\beta} \approx 0.623$$

For the CSSS collaboration network of participants, $\langle\cos\theta\rangle_{\text{real}} \approx \langle\cos\theta\rangle_{\text{random}}$, which indicates that participants choose their collaborators regardless of what genotypes they and their collaborators have, viz. there is no significant correlation pattern of genotypes between participants in the network. It is noteworthy to point out that a randomly mixed network is close to a disassortative network [5], in which vertices prefer to mix with vertices of dissimilar types. Here in the context of CSSS, our results demonstrate that participants with different background are well mixed and structured into projects groups, i.e., they are quite open to work with people with diverse (even dissimilar) academic or research background, which is the original motivation of CSSS, viz. to promote transdisciplinary collaboration. Although for the entire network, there is in average no significant correlation between genotypes in participants, the mean cosine similarity measure *within* each project group does vary. We are pursuing further study on this.

Acknowledgements

We are grateful to Juniper Lovato and John Paul Gonzales for hospitality at Santa Fe during the Complex Systems Summer School. Thanks to all of our CSSS 2013 classmates, whose collective intellectual participation motivated the emergence of this study. In particular, we are indebted to Johannes Schmidt and Stephan Lehner for keeping our project moving forward. Last but not least, we would like to give special thanks to Tom Carter, whose enlightening conversation (and "Dark Force") is the best memory of our midsummer night's dream at Santa Fe.

References

[1] SFI 2013 Complex Systems Summer School (CSSS 2013) project presentations website: http://tuvalu.santafe.edu/events/workshops/index.php/Presentations_2013.

[2] During the summer school at Santa Fe, we performed a survey (via personal conversation) on the academic & research background of CSSS 2013 participants.

[3] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, 2010.

[4] R. Guimerá, S. Mossa, A. Turtschi, and L. A. N. Amaral, *The worldwide air transportation network: Anomalous centrality, community structure and cities' global roles*, PNAS **102**, 7794-7799 (2005).

[5] M. E. J. Newman, *Mixing patterns in networks*, Phys. Rev. E **67**, 026126 (2003).