Novel Statistical Network Methodology to Identify and Analyze Cancer Biomarkers

Thomas E. Bartlett [†‡*] Sofia C. Olhede [‡] Alexey Zaikin [†§¶]

Abstract

The global burden of cancer is expected to increase between 2008 and 2030 from 13 to 22 million new cases each year (Bray et al. 2012). The strain that this will put on health services around the world will be reduced by the development of better methods for early detection of disease risk and progression.

The epigenome is thought of as the interface between the genome and the environment; hence measurements of DNA methylation, an epigenetic pattern, can indicate exposure to environmental risk factors. However these measurements are extremely noisy, making it a challenge to derive meaningful statistics from such data.

Using canonical correlation analysis we have developed a novel statistical measure of the level of interaction between a pair of genes (network nodes) in a single sample/patient, based on DNA methylation data. Testing this interaction measure for association with patient outcome, we show how to construct prognostic networks for cancer, in which the presence of a network edge indicates that the network interaction between the corresponding pair of genes (nodes) is statistically significantly prognostic. Detecting community structure in these networks by fitting the stochastic blockmodel allows novel cancer biomarkers to be detected. These findings represent new statistical tools for use in the biomedical sciences.

Key Words: Random network models, Network community detection, Multivariate statistics, Cancer biomarkers, Survival analysis, Epigenetics

1. Introduction

Complex systems which can be modelled as networks are ubiquitous. Well-known examples include social and economic networks, as well as many others in cell biology such as gene regulatory, metabolic and protein signalling networks. Over the past few years in cell biology, much of the focus has shifted from investigation of individual genes, to pathways of genes, to gene networks. The need for novel methodology for network analysis in cell biology results from this recognition that examining the way genes work in groups is often more successful in revealing biological principles. Further, by considering groups of genes together, statistical significance can be obtained which would not be possible at the level of individual genes.

The epigenome is thought of as the interface between the genome and the environment; hence, measurements of DNA methylation, an epigenetic pattern, can indicate exposure to environmental risk factors (Feinberg, Ohlsson, and Henikoff 2006; Cooney 2007). It is well established that DNA methylation plays a major role in gene regulation, and hence DNA methylation patterns often reflect patterns of gene regulation. Changes in DNA methylation are highly stochastic; however, the time-scale over which these changes take place (much

[†]Department of Mathematics, University College London, Gower St, London WC1E 6BT, United Kingdom

[‡]Department of Statistical Science, University College London, 1-19 Torrington Place, London WC1E 7HB, United Kingdom

[*]Email: thomas.bartlett.10@ucl.ac.uk

[§]Institute for Women's Health, University College London, 74 Huntley Street, London WC1E 6AU, United Kingdom

[¶]Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod 603950, Russia

quicker than mutations in the basic DNA code, but much slower than the varying expression level of individual genes), means that DNA methylation patterns are very promising for biomarker development. Further, there is much considerable evidence that changes in DNA methylation patterns are amongst the earliest genetic changes in oncogenesis (Feinberg, Ohlsson, and Henikoff 2006). DNA methylation data are extremely noisy; however, statistics which summarise DNA methylation patterns at the gene level show much promise as a way to represent and analyse these measurements (Bartlett et al. 2013).

Statistical network models are an efficient way to represent and analyse large numbers of variables and samples, over the very large quantities of data being produced by the latest technologies in cell biology, i.e., 'Big Data' (National Research Council 2013; Boulton et al. 2012; European Commission 2014). One such model which has received much attention is the 'stochastic blockmodel' (Holland, Laskey, and Leinhardt 1983; Bickel and Chen 2009), under which there is a greater probability of observing an edge (or interaction) between a pair of nodes if they are in the same block, or community. The Newman-Girvan modularity (Newman and Girvan 2004) quantifies the extent to which edges are observed between community members, for a particular assignment of nodes to communities, compared to the expected number of edges between community members if there were no community structure present. It can be shown that fitting the stochastic blockmodel is equivalent to maximising the Newman-Girvan modularity over a network, and that these are both equivalent to spectral clustering (Riolo and Newman 2012; Newman 2013).

The problem of finding communities in biological networks has been studied for many years (Girvan and Newman 2002), with communities in biological networks representing subnetwork modules with specific physiological functions. It has been shown recently that the stochastic blockmodel can be used to represent any network as a 'network histogram', whatever the generating mechanism of that network, as long as that network is not too sparse (Olhede and Wolfe 2014). Further, those advances provide a heuristic method to estimate the maximum number of blocks, or clusters, which a valid blockmodel representation of the network may contain. This is important and useful, because it means that the blockmodel can be used to identify an unknown number of communities, or functional subnetwork modules, in biological networks. Biological networks are known to display multi-scale properties (Barabási and Oltvai 2004; Palla, Lovász, and Vicsek 2010), which means that different functional organisation is visible at different granularities, or scales. Hence, the network histogram method (Olhede and Wolfe 2014) can be used to estimate the optimal granularity at which communities, or functional subnetwork modules, can be identified and isolated in biological networks, by fitting the stochastic blockmodel.

We have developed a novel statistical measure of the level of interaction between a pair of genes (network nodes) in a single sample/patient, based on DNA methylation data, which we term the 'DNA methylation network interaction measure' (Bartlett, Olhede, and Zaikin 2014). By testing this interaction measure for association with patient survival outcome, we show how to construct a binary prognostic network, in which the presence of a network edge indicates that the corresponding gene-gene interaction is statistically significantly prognostic. Community structure can be detected in these networks, by fitting the stochastic blockmodel; each community, or subnetwork module, identified in this way, represents a potential network biomarker. These findings represent new statistical tools for use in the biomedical sciences.

2. Methods and Models

2.1 DNA Methylation Network Interaction Measure

Using canonical correlation analysis (CCA) (Hotelling 1936), we have developed a novel statistical measure (Bartlett, Olhede, and Zaikin 2014), of the level of interaction between a pair of genes (network nodes) in a single sample/patient, based on DNA methylation data (figure 1). This DNA methylation network interaction measure quantifies the extent to which the DNA methylation profiles of a pair of genes explain each other. It is based only on measurements of the DNA methylation profiles of this pair of genes, and it acts as a surrogate for a measure of the extent to which this pair of genes behave interactively. Such interactive behaviour may include transcriptional regulation or other types of biochemical interaction, influencing gene expression levels and the presence of alternatively spliced gene products (Jones 2012), amongst other phenomena.

The DNA methylation network interaction measure is defined by analogy to CCA. CCA aims to discover linear combinations of variables of one type, and linear combinations of variables of another type, so that these combinations best explain each other. In this context, a particular way of combining (by scaling and adding) the deviations from the mean methylation profile at a number of locations within one gene might be particularly effective at explaining a particular combination (again, by scaling and adding) of the deviations from the mean methylation profile at a number of locations in another gene, and *vice-versa*. There will probably be fewer ways in which the methylation levels of these genes covary across the samples, than there are locations at which methylation is measured along the genes; this is because the methylation level is highly correlated at many locations along a particular gene. CCA finds the most important components of this covariation across samples.

CCA seeks to find the vectors a and b, in the p and q dimensional spaces of variables $\mathbf{X} = (x_1, x_2, ..., x_p)'$ and $\mathbf{Y} = (y_1, y_2, ..., y_q)'$ respectively, which maximise the correlation $\rho = \mathrm{cor}\,(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y})$, defined according to equation 1,

$$\rho = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{XY}}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}}} \tag{1}$$

where $\boldsymbol{\Sigma}_{\mathbf{XX}} = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})'\right]$ and $\boldsymbol{\Sigma}_{\mathbf{YY}} = \mathbb{E}\left[(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})'\right]$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ respectively, and $\boldsymbol{\Sigma}_{\mathbf{XY}} = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})'\right]$ is the cross-covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$.

Two genes X and Y have corresponding methylation profiles which are measured for sample / patient j at p and q CpGs (loci) respectively along these genes. Denoting these measurements by the variables $x_1, ...x_p$ and $y_1, ..., y_q$ for genes X and Y respectively, the DNA methylation profiles for these genes, for patient j, can be represented by the vectors $\mathbf{x}(j)$ and $\mathbf{y}(j)$, which have p and q entries respectively. A measure of DNA methylation network interaction $\rho_{XY}(j)$, of the methylation profiles of genes X and Y for sample j, can then be defined by analogy with equation 1, according to equation 2,

$$\hat{\rho}_{XY}(j) = \frac{\mathbf{x}(j)^T\hat{\boldsymbol{\Sigma}}^{(h)}_{XY}\mathbf{y}(j)}{\sqrt{\mathbf{x}(j)^T\hat{\boldsymbol{\Sigma}}^{(h)}_{XX}\mathbf{x}(j)}\sqrt{\mathbf{y}(j)^T\hat{\boldsymbol{\Sigma}}^{(h)}_{YY}\mathbf{y}(j)}} \tag{2}$$

where $\hat{\mathbf{\Sigma}}^{(h)}_{XX}$, $\hat{\mathbf{\Sigma}}^{(h)}_{YY}$ and $\hat{\mathbf{\Sigma}}^{(h)}_{XY}$ are estimated from healthy rather than cancer samples in the methylation data set, according to equations 3 - 5,

$$\hat{\mathbf{\Sigma}}^{(h)}_{XX} = \frac{1}{n_h} \sum_{j \in \text{healthy}} \left(\mathbf{x}(j) - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}}\right) \left(\mathbf{x}(j) - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}}\right)^T \tag{3}$$

$$\hat{\mathbf{\Sigma}}^{(h)}_{YY} = \frac{1}{n_h} \sum_{j \in \text{healthy}} \left(\mathbf{y}(j) - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{Y}}\right) \left(\mathbf{y}(j) - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{Y}}\right)^T \tag{4}$$

$$\hat{\mathbf{\Sigma}}^{(h)}_{XY} = \frac{1}{n_h} \sum_{j \in \text{healthy}} \left(\mathbf{x}(j) - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}}\right) \left(\mathbf{y}(j) - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{Y}}\right)^T, \tag{5}$$

where

$$\hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}} = \frac{1}{n_h} \sum_{j \in \text{healthy}} \mathbf{x}(j),$$

and

$$\hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{Y}} = \frac{1}{n_h} \sum_{j \in \text{healthy}} \mathbf{y}(j),$$

and n_h is the number of healthy samples in the data set. When the DNA methylation network interaction measure $\rho_{XY}(j)$ is large (i.e., close to 1), the corresponding pair of genes explain each other's transcriptional or translational behaviour (as reflected in their methylation profiles) well, or have otherwise well-correlated interactive behaviour, for sample/patient j. Hence, $\rho_{XY}(j)$ measures (according to their DNA methylation profiles) the level of interaction between genes X and Y in tumour sample j, compared to typical interactions between these genes in healthy tissue.

2.2 Prognostic Network Construction

We construct a prognostic interaction network for m genes, represented by the $m \times m$ adjacency matrix $\mathbf{A}$, by defining an edge to be present (i.e., $A_{ij} = 1$) if and only if the corresponding pair of genes (nodes) are significantly prognostic according to the DNA methylation network interaction measure (i.e., otherwise $A_{ij} = 0$). To do this, for each of the $\binom{m}{2}$ pairs of genes in the network, we use the Cox proportional hazards model (Cox 1972) to test the association of the DNA methylation network interaction measure $\rho_{XY}(j)$ for the pair of genes X and Y with patient survival outcome, across patients $j = 1, ...n$, adjusting for significant clinical covariates (to detect novel DNA methylation biomarkers which are independent of known prognostic clinical features).

We define an interaction as being statistically significant with $q < 0.1$, with respect to either of the interacting genes, where q is the false discovery rate (FDR) corrected (Benjamini and Hochberg 1995) p-value. This threshold FDR is set relatively high at 0.1, because we expect significantly prognostic interactions, i.e., true positives, to tend to group together in subnetwork modules/communities, and false positives not to do so; hence we expect the false discovery rate to be much lower than 0.1 in practice after community detection, because only a fraction of the $\binom{m}{2}$ possible edges occur between pairs of nodes which are in the same community. In a simulation study (data not shown), we found that applying this methodology to a network of randomly generated p-values (from a uniform distribution on $[0, 1]$), no community structure is detected beyond single-figure groups of nodes. Therefore, we set 10 nodes as the lower limit of the allowable size of a detected subnetwork module, as a potential prognostic network biomarker. In setting this FDR threshold, we also verified that the degree distributions of the resulting networks display the power law behaviour expected in biological networks (Wagner 2002; Barabási and Oltvai 2004).

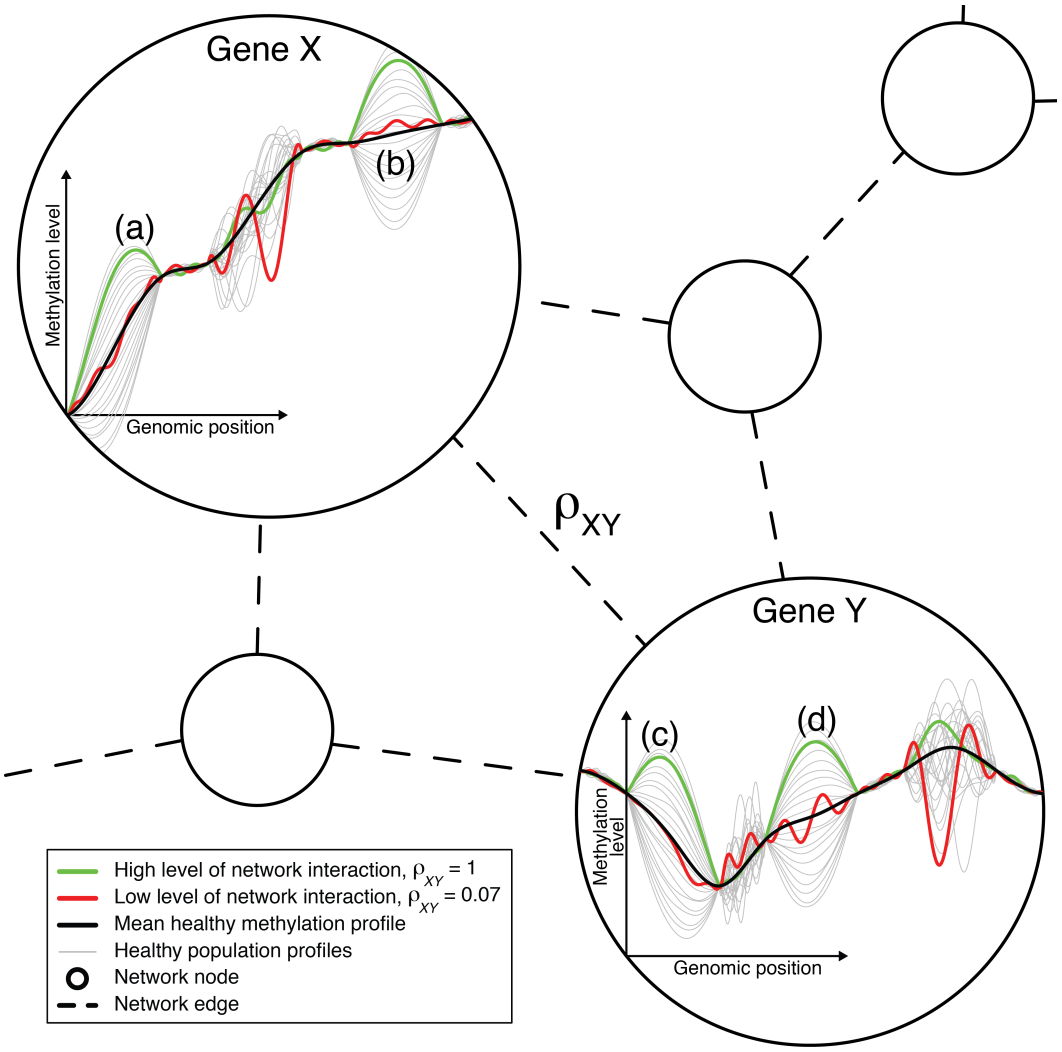

Figure 1: **The DNA methylation Network Interaction Measure.** *A combination of the variation of the healthy methylation profiles in regions (a) and (b) of gene X explains well / is well-explained by a combination of the variation of the healthy methylation profiles in regions (c) and (d) of gene Y. The green cancer sample varies by a large amount about the mean methylation profile and in a typical way in these regions in both genes. Hence, the green sample corresponds to a high level of network interaction for this sample, $\rho_{XY} = 1$. The variation in the other regions of these genes do not well-explain each other, and so the red sample, which varies by a large amount in these other regions and varies less and in an atypical way in regions (a)-(d), corresponds to a low level of network interaction, $\rho_{XY} = 0.07$. Genes X and Y are likely to have different numbers of methylation measurement locations (i.e., variables X and Y are of different dimension). The ordering of the measurement locations has no influence on the calculation of ρ, as long as the ordering is consistent across samples.*

2.3 Community and Biomarker Detection

Network nodes can be grouped together, according to their propensity to interact with each other, for example groups of friends in a social network, or functional subnetwork modules in a biological network; this statistical method is often referred to as community detection (Girvan and Newman 2002; Newman 2004). Hence, community detection allows us to find groups of genes in our constructed prognostic network, i.e., prognostic subnetwork modules, which interact differently in cancer than in healthy tissue, in a way which is predictive of how advanced the disease is. Within such a detected community/subnetwork module,

the genes may interact with each other more (relative to healthy tissue) the more serious the disease is (as is predominantly the case in figure 2 (a)), or they may interact with each other less the more serious the disease is, or both these scenarios may arise within the same community/subnetwork module (as in figure 2 (b)). We carry out community detection by fitting the stochastic blockmodel (Holland, Laskey, and Leinhardt 1983; Bickel and Chen 2009), using the network histogram choice of number of blocks in a valid stochastic blockmodel of the network (Olhede and Wolfe 2014). Each community, or subnetwork module, identified in this way, represents a potential network biomarker. For each such biomarker, a prognostic measure can be calculated, by summarising the DNA methylation network interaction measure over the subnetwork module/community.

3. Results

Figure 2 shows examples of two network biomarkers, detected by our methodology in a breast cancer data set downloaded from *The Cancer Genome Atlas* (Hampton 2006; Collins and Barker 2007). The network biomarkers are detected in a group of $n = 273$ patient tumour samples (including 23 events), and validated in a separate, independent group of $n = 398$ samples (including 29 events). Validation p-values are 7.2×10^{-4} and 2.4×10^{-3} for the network biomarkers shown in figure 2 (a) and (b) respectively, with corresponding FDR q-values 0.029 and 0.049, respectively. Gene-set enrichment analysis (Subramanian et al. 2005) based on the gene-sets available for download from from The Broad Institute Molecular Signatures Database (*www.broadinstitute.org/gsea/msigdb*) shows significant enrichment of the network biomarker shown in figure 2 (a) by 12 gene-sets (FDR $q < 0.05$, including seven gene-sets with $q < 10^{-6}$) relating to stem cell genes, and known patterns of differential methylation in cancer. However, the network biomarker in figure 2 (b) does not show significant enrichment by any of the gene-sets in the Broad Institute Molecular Signatures Database, meaning that it is likely to represent a novel biological finding.

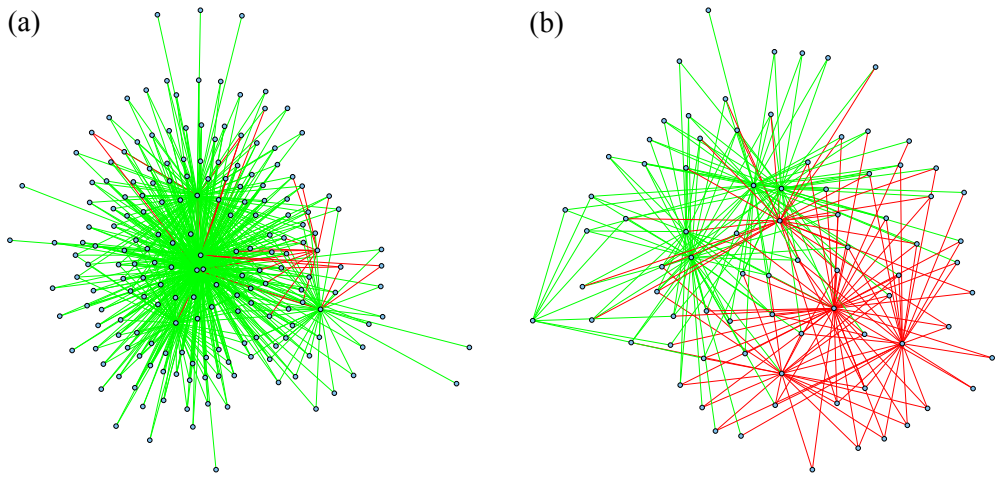

Figure 2: **Detected Network Biomarkers.** *Network biomarkers detected in a breast cancer data set, with (a) 191 and (b) 82 nodes. The green and red edges represent network interactions which increase and decrease (relative to healthy tissue) with worse patient prognosis, respectively; hence, these are examples of network re-wiring in cancerous tissue.*

4. Discussion

Our new methodology enables the discovery of DNA-based prognostic biomarkers, as sub-network modules, or communities, in a prognostic network constructed from gene-gene interactions in DNA methylation data. These findings represent new statistical tools for use in the biomedical sciences, and we hope that they will ultimately aid the development of new prognostic biomarkers of real clinical value.

5. Acknowledgements

TEB acknowledges support by the UK Engineering and Physical Sciences Research Council (ESPRC) and the UK Medical Research Council (MRC), via UCL CoMPLEX. SCO acknowledges funding from EPSRC grant no. EP/I005250/1. AZ acknowledges support from the Russian Foundation for Basic Research (14-02-01202). We are grateful to all specimen donors and research groups involved in providing data used in this study via TCGA.

REFERENCES

Barabási, A.-L. and Oltvai, Z. N. (2004). "Network biology: understanding the cell's functional organization". *Nature Reviews Genetics* 5.2, 101–113.

Bartlett, T. E., Olhede, S. C., and Zaikin, A. (2014). "A DNA Methylation Network Interaction Measure, and Detection of Network Oncomarkers". *PloS one* 9.1, e84573.

Bartlett, T. E. et al. (2013). "Corruption of the Intra-Gene DNA Methylation Architecture Is a Hallmark of Cancer". *PloS one* 8.7, e68285.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bickel, P. J. and Chen, A. (2009). "A nonparametric view of network models and Newman–Girvan and other modularities". *Proceedings of the National Academy of Sciences* 106.50, 21068–21073.

Boulton, R et al. (2012). "Science as an open enterprise". *Royal Society, London* 104.

Bray, F. et al. (2012). "Global cancer transitions according to the Human Development Index (2008–2030): a population-based study". *The lancet oncology* 13.8, 790–801.

Collins, F. and Barker, A. (2007). "Mapping the cancer genome". *Scientific American Magazine* 296.3, 50–57.

Cooney, C. A. (2007). "Epigenetics-DNA-based mirror of our environment?" *Disease Markers* 23.1, 121–137.

Cox, D. R. (1972). "Regression models and life tables (with discussion)". *Journal of the Royal Statistical Society* 34, 187–220.

European Commission (2014). *Commission urges governments to embrace potential of Big Data*. URL: http://europa.eu/rapid/press-release_IP-14-769_en.htm.

Feinberg, A., Ohlsson, R., and Henikoff, S. (2006). "The epigenetic progenitor origin of human cancer". *Nature reviews genetics* 7.1, 21–33.

Girvan, M. and Newman, M. E. (2002). "Community structure in social and biological networks". *Proceedings of the National Academy of Sciences* 99.12, 7821–7826.

Hampton, T. (2006). "Cancer genome atlas". *JAMA: The Journal of the American Medical Association* 296.16, 1958–1958.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). "Stochastic blockmodels: First steps". *Social networks* 5.2, 109–137.

Hotelling, H. (1936). "Relations between two sets of variates". *Biometrika* 28.3/4, 321–377.

Jones, P. (2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond". *Nature Reviews Genetics* 13.7, 484–492.

National Research Council, C. C. D. (2013). *Frontiers in Massive Data Analysis*. The National Academies Press. ISBN: 9780309287784. URL: http://www.nap.edu/openbook.php?record_id=18374.

Newman, M. E. (2004). "Detecting community structure in networks". *The European Physical Journal B-Condensed Matter and Complex Systems* 38.2, 321–330.

Newman, M. E. and Girvan, M. (2004). "Finding and evaluating community structure in networks". *Physical review E* 69.2, 026113.

Newman, M. (2013). "Spectral methods for network community detection and graph partitioning". *arXiv preprint arXiv:1307.7729*.

Olhede, S. and Wolfe, P. (2014). "Network histograms and universality of blockmodel approximation". *Proceedings of the National Academy of Sciences*, In press.

Palla, G., Lovász, L., and Vicsek, T. (2010). "Multifractal network generator". *Proceedings of the National Academy of Sciences* 107.17, 7640–7645.

Riolo, M. A. and Newman, M. (2012). "First-principles multiway spectral partitioning of graphs". *arXiv preprint arXiv:1209.5969*.

Subramanian, A. et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". *Proceedings of the National Academy of Sciences of the United States of America* 102.43, 15545.

Wagner, A. (2002). "Estimating coarse gene network structure from large-scale gene perturbation data". *Genome research* 12.2, 309–315.