TEMPLE UNIVERSITY

# Network clustering of cancer patients based on DNA methylation variability

CIS 5524: Analysis and Modeling of Social and Information Networks

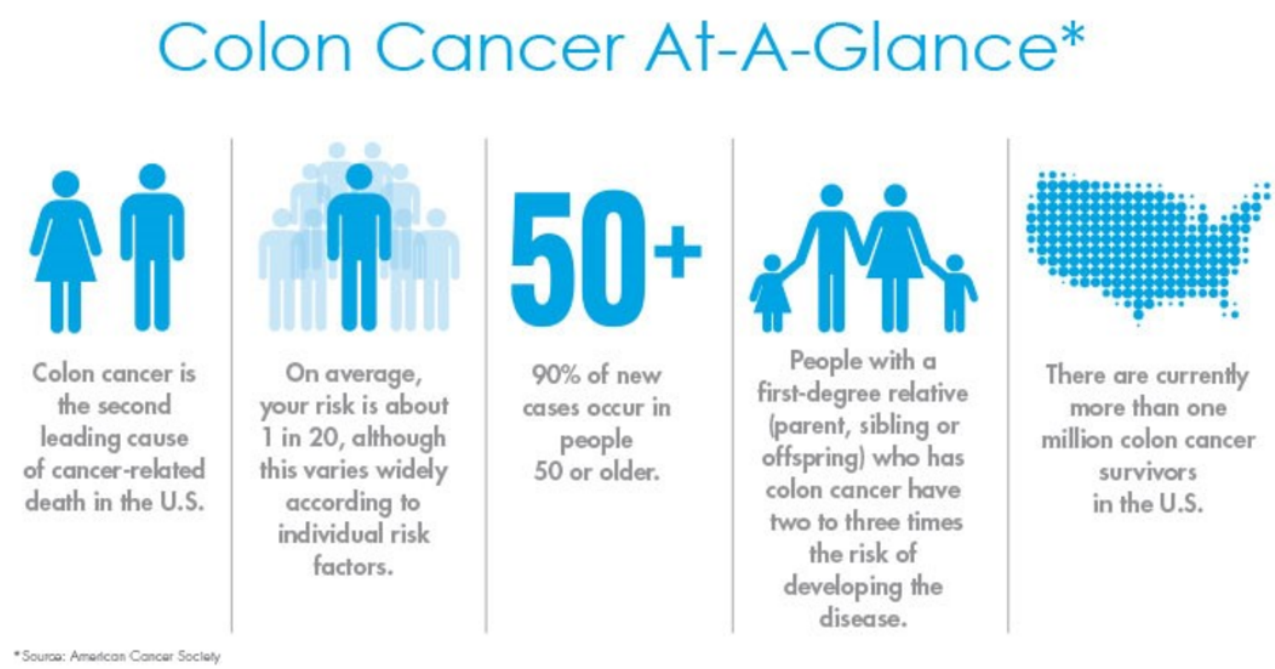
Marija Stanojevic

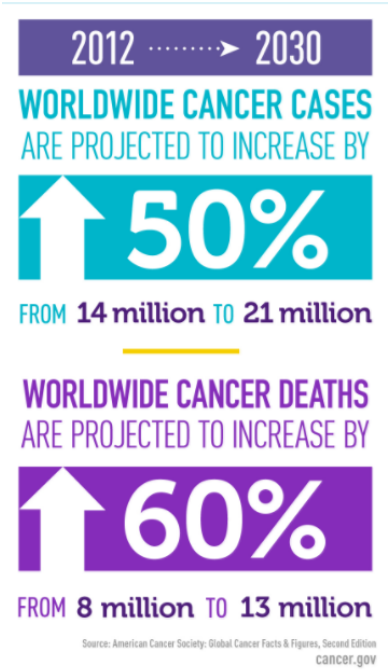
marija.stanojevic@temple.edu

Department of Computer and Information Sciences

27th April 2017

# Introduction




**Background and motivation:**

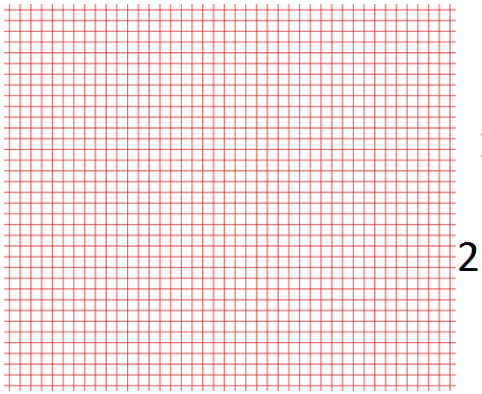
- Methylation influenced by genetics and environment/behavior
- Experimental observations indicated that certain people are outliers in many significant methylation points
- **Hypothesis**: outliers are caused by genetic mutations. This paper aim to check if experimental observations are valid and to group those people

**Objective:**

Cluster health/ cancerous tissues in groups based on similarity of significant features in their methylation arrays

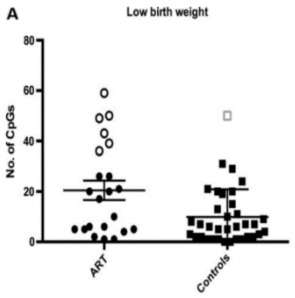
# Data and preprocessing

- **Data is taken from TCGA project for colon cancer, using GDC Data Portal API**
- **Downloaded files:**
  - Methylation Beta Value (458 cases, 556 files)
  - Biospecimen Supplement (461 cases), contains information about samples
  - Clinical Supplement (459 cases), contains information about patients
- **Only 75 patients have data from health tissues, so totally 150 samples data are considered: 75 from health and 75 from cancer tissues for same patients**
- **Methylation files contained 485578 or 27579 positions, intersection is taken**
  - Methylation positions from sex chromosomes X and Y are removed
  - Methylations with more than 20% of missing values are removed
  - Methylations with less than 20% of missing values are imputed with MEAN
- **Total number of methylations / sample = 22385**
  - Mean, stddev are calculated for each methylation position and deviation is calculated for each patient and methylation position



# Methodology

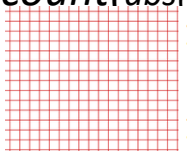
**Feature selection: get features with high variability**

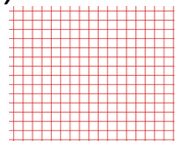



***1.ST**: stddev > threshold*

***2.SMT**: stddev/mean > thresh*

***3.D**: count(abs(deviation) > 2 stddev) > threshold*


150 x 100


150 x 2000

**Clustering: normalization of correlation values with next metrics:** $\frac{1+corr}{2}$ **and** $abs(corr)$

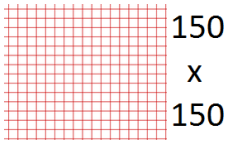
*1.Louvain Modularity*

$$Q = \frac{1}{2m}\Sigma_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

*2.Spectral Clustering (eigenvalues and eigenvectors of Laplacian matrix)*

**Correlation: results in complete graph of patients**


150 x 150

*1.Pearson correlation (linear relationship)*

$$\rho_{X,Y} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

*2.Spearman correlation (monotonic relationship)*

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

*3.Kendall correlation (ordinal association)*

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

**Evaluation:**

*1.Modularity*

$$Q = \frac{1}{2m}\Sigma_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

*2.Conductance*

$$\phi(G) = avg(\phi(C_i)) \ , \ C_i \subseteq V$$

$$\phi(C_i) = \frac{\sum_{u \in C_i}\sum_{v \notin C_i} w(\{u, v\})}{min(a(C_i), a(\bar{C_i}))}$$

*3.Coverage*

$$coverage(C) = \frac{w(C)}{w(G)}$$

$$w(C) = \sum_{i=1}^{k} w(E(v_x, v_y)); \ v_x, v_y \in C_i$$

# Results

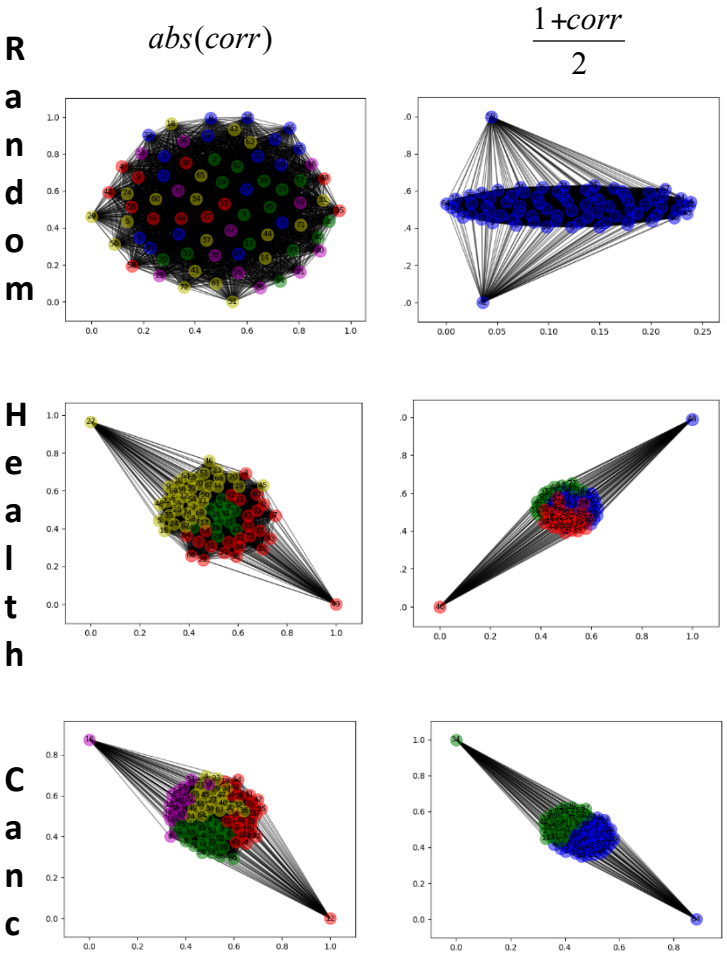
**Null model: random features permutation**

- *Much more clusters or one cluster*
- *Networks show no groups*

**Results real data**

*1.2-4 clusters as expected*

*2.Clusters are meaningful*




| Clustering + Evaluation Method | Random $abs(corr)$ | Random $\frac{1+corr}{2}$ | Health $abs(corr)$ | Health $\frac{1+corr}{2}$ | Cancer $abs(corr)$ | Cancer $\frac{1+corr}{2}$ |
|---|---|---|---|---|---|---|
| **Modularity (Louvain)** | 0.099249, Pearson, 100, SMT | 0.005466 Pearson, 100, D | 0.162885 Kendall, 100, ST | 0.065379 Pearson, 2000, SMT | 0.204630 Kendall, 100, D | 0.243234 Spearman, 100, SMT |
| **Conductance (Louvain)** | 4.745966 Spearman, 2000, D | 1.982202 Pearson, 100, SMT | 1.393586 Kendall, 100, SMT | 1.622324 Pearson, 2000, ST | 1.091775 Kendall, 2000, ST | 0.690853 Spearman, 100, SMT |
| **Coverage (Louvain)** | 0.348093 Kendall, 2000, D | 1 – single cluster Multiple | 0.642946 Kendall, 2000, D | 0.661576 Spearman, 2000, SMT | 0.694147 Kendall, 2000, ST | 0.743313 Spearman, 100, SMT |
| **Modularity (Spectral)** | 0.075320 Pearson, 100, SMT | 0.004795 Pearson, 100, D | 0.149705, Kendall, 100, ST | 0.064327 Pearson, 2000, SMT | 0.202401 Kendall, 100, D | 0.243234 Spearman, 100, SMT |
| **Conductance (Spectral)** | 4.485229 Kendall, 2000, D | 1.978286 Pearson, 100, ST | 1.393586 Kendall, 100, SMT | 1.657455 Pearson, 2000, SMT | 1.057027 Kendall, 2000, ST | 0.690853 Spearman, 100, SMT |
| **Coverage (Spectral)** | 0.436908 Pearson, 100, SMT | 1 – single cluster Multiple | 0.593260 Kendall, 100, SMT | 0.744199 Spearman, 2000, SMT | 0.66919 Spearman, 2000, ST | 0.743313 Spearman, 100, SMT |

* cells contain best value received from 100 or 2000 selected features with one of the ST, SMT or D feature selection methods and with one of the three correlations (Pearson, Spearman, Kendall)

** Modularity [-1, 1] and coverage [0, 1] is better if higher, conductance [0, +∞] is better if lower

# Conclusion

- **Clusters are meaningful. Clustering is better for cancer samples data than for health samples data. As expected there are 2-4 clusters**
- **$\frac{1+corr}{2}$ metrics gives always better results for cancer tissue data and for health tissue data under coverage evaluation, but $abs(corr)$ gives better results for health tissue data for other evaluations**
- **SMT feature selection gives best results in 19/36 cases**
- **All correlation methodologies are equally represented in best results**
- **Both clustering methods give similar results under all evaluation metrics**
- **Evaluation methods are complemental (show different aspects of clustering)**
- **Future work:**
    - Analyze overlapping of features under different selection methods
    - Examine which patients are always in same clusters
    - Study if those patients have common genetic features
    - Develop multi-level networks clustering model that will give better clustering results (different feature selections and correlation methods)