Vietnam Journal of Science and Technology ***60*** *(6) (2022) 1134-1148*

doi:10.15625/2525-2518/16678

IDENTIFICATION OF CANCER RULES IN VIET NAM BY NETWORK MODULARITY

Minh Tan Nguyen[1], Duc Tinh Pham[2], Viet Ha Tran[3], Tien-Dzung Tran[3, *]

[1]*Center of Information – Library, Hanoi University of Industry, 298 Cau Dien Street,*
Bac Tu Liem District, Ha Noi, Viet Nam

[2]*Graduate University of Science and Technology, Vietnam Academy of Science and Technology,*
18 Hoang Quoc Viet, Cau Giay, Ha Noi, Viet Nam

[3]*Department of Software Engineering, Faculty of Information Technology, Hanoi University of Industry, 298 Cau Dien Street, Bac Tu Liem District, Ha Noi, Viet Nam*

[1,2,3]*Complex Systems and Bioinformatics Lab, Hanoi University of Industry,*
298 Cau Dien Street, Bac Tu Liem District, Ha Noi, Viet Nam

[*]Email: *trantd@haui.edu.vn*

Received: 2 December 2021; Accepted for publication: 19 March 2022

Abstract. Data clustering tools can uncover new knowledge to be used in cancer diagnosis and treatment. In this study, we proposed a novel method to cluster records of a relation. First, we designed an algorithm that calculates the similarity between record pairs of the relation, and then this similarity measure was used to generate a network corresponding to the relation. Finally, we used a Network science technique to detect clusters of records from the network and extract insights from the clusters. Applying the method to mine a cancer-screening dataset at the Vietnam Central Cancer Hospital with over 177,000 records, we have discovered several new cancer laws in Viet Nam, which contribute to cancer detection and treatment support. It is disclosed from these cancer rules that some types of cancer run in certain family lines and living places in Viet Nam. Clustering a relation by Network science approach can be a good choice for mining large-scale relational data.

Keywords: network modularity, cancer rule identification, network inference, graph mining

Classification numbers: 4.8.5

1. INTRODUCTION

As a country in Southeast Asia and having a diverse human genome, Viet Nam is in the third group of countries with a relatively high rate of cancer and mortality, ranking 91 out of 185 countries and territories with an ASR (age-standardised rate) of 159.7 / 100,000 people according to the International Agency for Research on Cancer (2020) [1, 2]. Along with the growth of cancer globally, the number of cancer patients in Viet Nam has also increased exponentially with all ages, occupations, and social classes. According to many researchers, the cancer rate in Viet Nam will increase significantly in the coming years [3 - 6]. Cancer is one of the most dangerous diseases, directly affecting life and standing only behind cardiovascular diseases [7 - 9]. The rules of cancer are very diverse, possibly due to the genetic or environmental factors [10 - 12]. Indeed, several recent studies have shown that certain biomarker genes are responsible for the origin of many cancers in the presence of genetic mutations [13, 14]. During an analysis process of cancer cell signaling pathways, it was found that if a lung cell had a mutation of the KRAS, EGFR, or ERBB2 genes, it would lead to non-small cell lung cancer. Moreover, the COL4A1 gene mutation is often found in small-cell lung cancer [15]. Besides genetics, environmental factors such as tobacco smoke, alcohol [16 - 19], water pollution [20 - 22], unsafe food [23, 24], and air pollution [25, 26] have also been fully reported as cancer cause. In addition, bad living habits sometimes have a bad effect [27, 28]. For example, a higher risk of stomach cancer is often related to a diet with high sodium nitrate [29, 30]. Likewise, certain types of digestive system cancers can be caused by consuming inappropriate foods [31 - 33]. To contribute to the fight against cancer, traditional data analysis tools such as SAS, SPSS, STATA, and R are often used to extract and retrieve insights from cancer datasets [2, 34 - 36]. One of the most popular analytical methods integrated inside these tools for data mining is clustering algorithms. Determining a suitable clustering algorithm will directly affect the clustering results. K-means is a famous data-clustering algorithm that is often deployed in various cases [37 - 40]. However, the standard K-means has certain shortcomings, such as the requirement to determine in advance the number of K-clusters [41] and to compute the distance between objects of each data as well as all cluster centers in each iteration, which result in low levels of clustering efficiency [42].

Table 1. Cancer screening data.

ID	Family	Age	Sex	Address	Top	Diagnosis	Conclusion
1631	Le	41	1	Thanh Hoa	C50.9	K Breasts	Infiltrative papillary carcinoma. Ipsilateral axillary lymph node hyperplasia (8 lymph nodes)
1625	Nguyen	46	1	Ha Nam	C53.9	polyp	Chronic inflammation of the cervix. Fibrocystic polyps
1630	Nhu	74	2	Ha Noi	C77.9	Tuberculosis	Necrotizing inflammatory organization
1637	Nguyen	36	1	Quang Tri	C02.9	K	Infiltrative squamous cell carcinoma, grade IV. Cervical lymph node metastasis

In this study, we show a network approach clustering method that overcomes the limitations of the algorithms used hitherto for clustering cancer screening datasets. We exploited a 7-attribute relation containing cancer-screening test information at K-Hospital, Viet Nam. After pre-processing this relation, we built a weighted undirected graph (network), where each node was a tuple, and each edge and weight represent the similarity between two tuples. Two nodes were connected if their similarity is not below than a given threshold α where α was determined by experimental observation. Applying the maximum network modularity algorithm on the graph, we discovered network modules that were so-called clusters which share common values of attributes. From these clusters, we extracted novel cancer laws, providing important support for the prevention, detection, and treatment of cancer patients in Viet Nam.

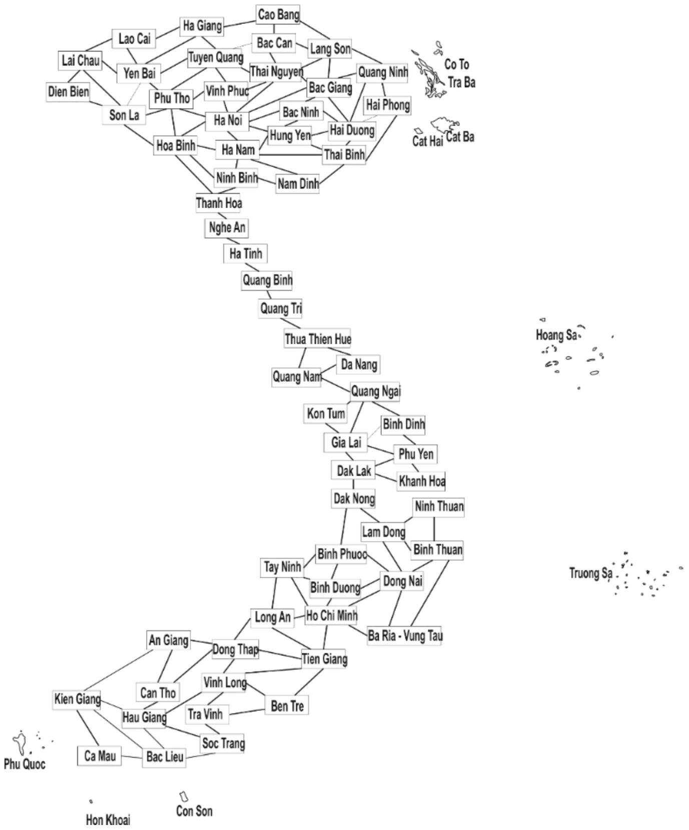

Figure 1. Network model of Viet Nam map: the graph has 66 nodes (63 provinces and 12 district islands) and 128 edges, where each node represents a province/city and each edge indicates the border between two provinces.

2. MATERIALS AND METHODS

2.1. Cancer screening data

The dataset was gathered from the record-keeping software about patients who took treatment at K-Hospital (Vietnam National Cancer Hospital) in Ha Noi from February 2009 to June 2014. After analyzing and synthesizing the data system, initial relation R1 with 177,565 patient records and 15 attributes was obtained. Due to the patient's privacy, the patient's given name was trimmed. The analysis removed invalid data sets and attributes for 27 northern provinces of Viet Nam and normalized three attributes containing text data types and inconsistent information such as Last name, Address, and Conclusion, and the intermediate relation R2 (see Table 1) was obtained among 122,379 records, described by 07 attributes. Next, we selected the set of conclusions containing the keywords {K, cancer, carcinoma, malignant, metastatic lymph nodes, melanoma} from the R2 dataset and obtained cancer relation R consisting of 43,629 modeled records. described by attributes such as ID, Family, Age, Sex, Address, Conclusion, Topological (Top) – location, and type of cancer in the patient's body. The Topological (Top) records were encrypted according to the structure CXY.z, where C stands for cancer, XY for cancer type, z is the id for the exact area of cancer (for example, C50.9 means malignant breast cancer), and the value of XY.z was taken.

2.2. Standardized text data

Text data type fields such as Address, Family, and Conclusion typically contain extra or special characters. To clean the data (the cleaner the data, the better the analysis), special characters such as “.”, “,” , “!”, @ must be removed from the relation. In addition, we removed the extra words (stop-words) in the sentence such as that, then, is, how, well, maybe, yes, yeah. These words were duplicated multiple times in the text and did not make sense. We implemented the text data normalization algorithm as follows:

Algorithm 1. Standardized text data

```
function [s] Standardize_Text(t) // t is the text string to normalize
begin
  s ← t
  convert s to lower letter
  for each string in the array s
      remove character in s by regex “[\\\”\\\”’’.,():;+-=|]”
  end for
  a ← read a file containing stop-words
  for each string in the array s
      remove words in array s, which appear in the array a
  end for
  return s // s is the standardized text string
end
```

2.3. Digitizing the conclusion field data pair into vector pair

Doctors’ conclusions regarding their patients’ conditions was described in the field Conclusion. Bag-of-Words (BoW) algorithm was utilized to digitize a pair of strings using their semantic similarity. BoW is an algorithm in natural language processing to classify text. The idea of BoW is to separate and group text according to the “Bag of Words”.

Algorithm 2. Digitize the conclusion field data pair into vector pair

```
function [a, b] BoW(Conclusion1, Conclusion2)
// Input: Conclusion1, Conclusion2 as two strings
// Output: Two corresponding vectors
begin
  A = words in Conclusion1;
  B = words in Conclusion2;
  C = A ∪ B; //Disjoint union
  Vector_a = the number of occurrences of each element of C in A;
  Vector_b = the number of occurrences of each element of C in B;
  return a, b;
end
```

Algorithm 3. Create a BFS matrix representing distance value between pairs of provinces.

```
function [BFS]  BFS(G(V,E))
// Input: G is the graph, V is the set of vertices, E is the set of edges
 //Output: Matrix $BFS_{nxn}$ to store the distance value between provinces
begin
   for (start←0; start<|V|; start++;)
         checkedList ← new Boolean();
         Set all nodes to "not visited"
         queue ← new LinkedList() ;
        linkList ← new LinkedList();
         checkedList[start] ← true;
         pair ← new Pair(start ,1);
         queue. enqueue(pair);
         while queue is not empty do
               pair ← queue. dequeue();
              BFS [start][pair.vertex] ← pair.value;
              i ← linkList[pair.vertex];
              while every edge (i,x) $\in$ E  do
                    if (x has not been approved) then
                          checkedList[x] ← true;
                          p ← new Pair(x, pair.value + 1);
                          queue. enqueue(p);
                    end if
              end while
          end while
    end for
    return BFS; //The BFS is a matrix that represents distance between provinces
end
```

With the new test data, we tried to find out how many occurrences of each word of the test data appears in the "bag". The algorithm is presented in Algorithm 2.

2.4. Building the distance matrix of provinces

Addresses represents the names of regions in a geographic map and are stored as a text-based data type. To vectorize this, we sorted the names of the places alphabetically and assigned a number between [1..n] based on the first letter of the place, where n is the number of places (Table S1 in supplementary file). Next, based on the principle that adjacent provinces are linked together, and each province is only connected once, we made a list of contiguous provinces according to geographical factors. For example, Phu Tho, Vinh Phuc, Thai Nguyen, Bac Giang, Bac Ninh, Hung yen, Ha Nam, and Hoa Binh provinces are adjacent to Ha Noi. Based on the data, the Cytoscape tool (cytoscape.org) was utilized to visualize the province-linked network model in Viet Nam, as displayed in Figure 1. Next, the Breadth-First-Search (BFS) was enhanced to traverse a network of n vertices/provinces by width, thus computing and assigning distance value between pairs of provinces (see Algorithm 3). Then we created a BFS_{nxn} symmetry matrix, in which the BFS[i, j] denotes the distance value between pairs of provinces "*i*" and "*j*". Specifically, the BFS_{63x63} matrix that indicates the distance value between pairs of provinces in Viet Nam was built by Algorithm 3.

3. RESULTS AND DISCUSSION

3.1. A novel network construction algorithm

In this section, a novel algorithm was proposed to compute the similarity between pair records of a relation *R*. After the process of encoding the pair of data records, the algorithm applied Euclidean measurements to calculate the distance between vector pairs. Euclidean measurement has been used commonly in network-clustering studies [43 - 47]. The details of the algorithm for computing the similarity are represented as follows:

Algorithm 4. Computation of the similarity of two records

```
function [similarity] Similarity(a, b ⊆ R)
//Input: Two records a,b⊆ R (A1, A2,..., An) where Ai (i = 1..n) is specified in one of the following data types: Label, Address, Text and Number.
// Output: Similarity of a pair of records a, b
begin
        $\vec{x} = [x_1, x_2, x_3, \ldots x_n]$
        $\vec{y} = [y_1, y_2, y_3 \ldots y_n]$
        $x_i = \begin{cases} 1, \text{ if typeof}(a[i]) \in \{Label, Address\} \\ a[i], \text{ if typeof}(a[i])\text{=Number} \\ BoW(a[i], b[i])[0], \text{ if typeof}(a[i])\text{=Text} \end{cases}$
        $y_i = \begin{cases} \begin{bmatrix} 1, \text{ if b}[i]\text{=a}[i] \\ 0, \text{ if b}[i] \neq a[i] \end{bmatrix}, \text{ typeof}(b[i])\text{=Label} \\ b[i], \text{ if typeof}(b[i])\text{=Number} \\ BFS(a[i], b[i]), \text{ if typeof}(b[i])\text{=Address} \\ BoW(a[i], b[i])[1], \text{ if typeof}(b[i])\text{=Text} \end{cases}$
    $d = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \ldots + (y_n - x_n)^2};$
    $similarity(a, b) = \frac{1}{1+d};$
  end if
  return similarity(a,b);
end
```

Example 1. We consider a pair of records as follows:

a = {nguyen, 39, 2, 24, 77.9, Squamous cancer invades degree ii};

b = {nguyen, 38, 1, 24, 77.9, Squamous cancer invades}

Applying the algorithm, we have similarity(*a*,*b*) ≈ 0.333

The algorithm that computes the similarity between two records was used to create the network from a relation as follows:

Algorithm 5. Network construction

```
function [G]  NetworkConstruction(R (A1, A2,..., An), α∈[0,1])
//Input: + R (A1, A2,..., An) is a relation;
//         + α is the threshold
//Output:  Relation with two attributes {Start, End}
 represents the network
Begin
```

```
G←new Relation(Start, End)
foreach a in R
    foreach b in R
        t ← Similarity(a,b)
        if(t≥α)
            G← G ∪ (a, b)
    end for
end for
return G
end
```

3.2. The network built from the cancer relation

In order to detect clusters in the network, Algorithm 5 was applied to the cancer relation R (ID, Family, Age, Sex, Address, Top, Conclusion), which includes 43,629 records. To identify the pairs with the highest or greatest similarities conforming to a particular criterion, a threshold α was selected. In this research, the threshold α was evaluated so that the clustering purpose was achieved according to the criteria ensuring that the scale-free characteristic of the network and pairs of records in each cluster overlapped at the level of 03 properties, equivalent to 50 % of the analyzed attributes.

From examining the α values of record pairs in Table 2, we found that a threshold of 0.333 is suitable for the study. In other words, we reduced the number of pairs of similar levels to [0.333; 1] and obtained a G graph with 19,966 nodes and 68,213 edges. This was visualized by the Cytoscape software (Figure 2).

3.3. The clustering technique using network modularity optimization

Regarding detection of record/object clusters, the above complex was created where a record was represented by each node, and an edge was connected to two nodes if their similarity was higher than a given threshold α, specified by experimental observation. After the network was formed, modules (clusters) within the network were detected by the network modularity described in [48]. The clusters include similar objects, and each cluster extracted each rule of cancer. The network modularity function is described as follows:

$$Q = \frac{1}{2m}\sum_{vw}[A_{vw} - \frac{k_v k_w}{2m}]\frac{S_v S_w + 1}{2} \quad (1)$$

Table 2. Investigating the similarity of records by α.

ID	Family	Age	Sex	Address	Top	A
607181	Mai	41	1	42	C50.9	1
607207	Mai	41	1	42	C50.9	
121054649	Nguyen	39	1	24	C53.9	0,5
121058256	Nguyen	39	1	62	C53.9	
679	Mai	50	1	60	C50.9	0,414
1186516	nguyen	50	1	60	C50.9	
679	mai	50	1	60	C50.9	0,333
656740	hoang	50	1	63	C50.9	

The case α = 0.333. Two records overlap at data level 03 attributes

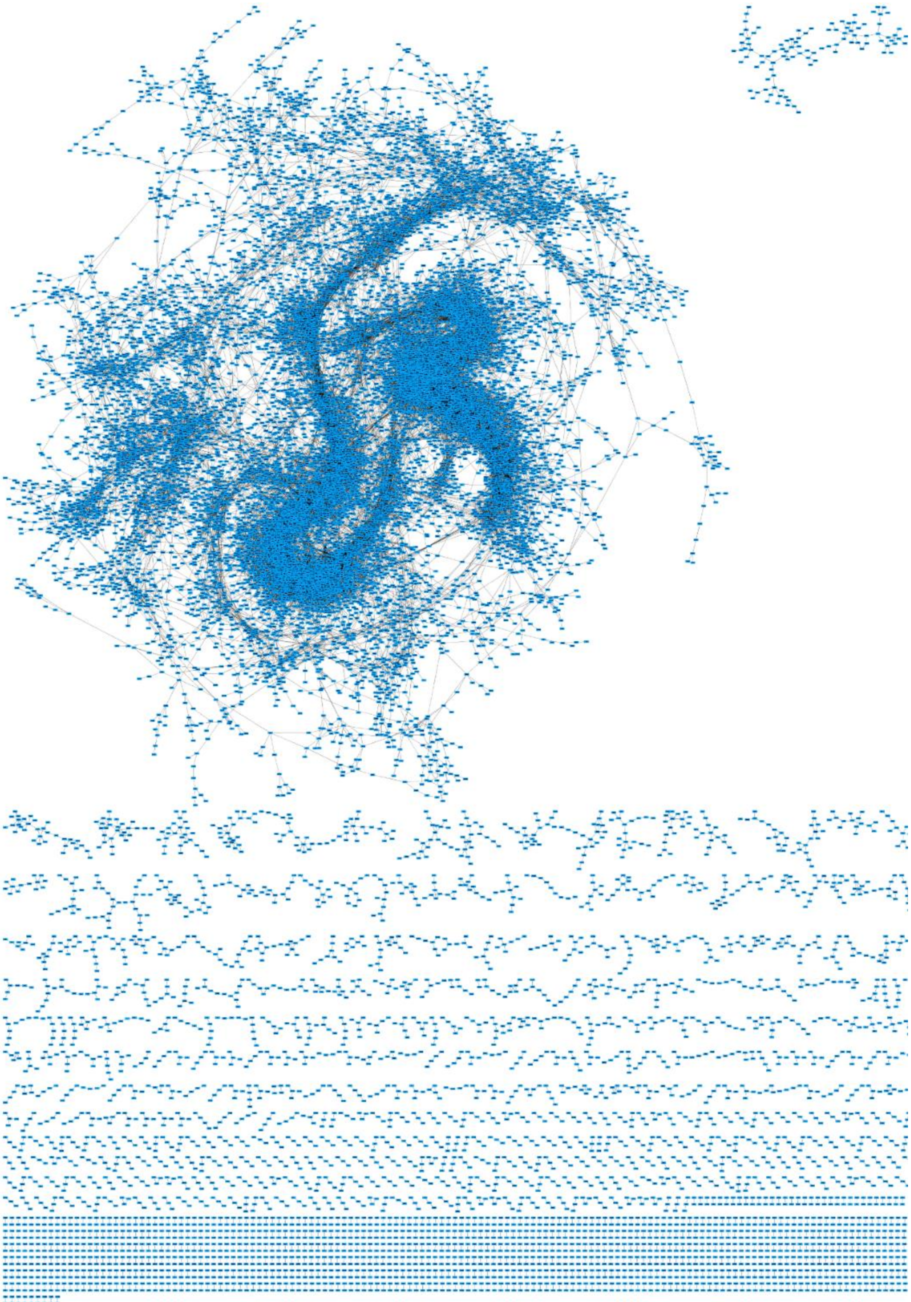

Figure 2. Network *G* was created corresponding to threshold $\alpha = 0.333$ and consisted of 19,966 nodes and 68,213 edges where its architecture shows the scale-free property. The threshold was chosen from the investigation illustrated in Table 2.

Assume: m is the number of links, n is the number of nodes, k_v and k_w are the number of in-/out- edges of nodes v and w; $A_{vw} = 0$ if there is no edge between v and w, otherwise $A_{vw} = 1$ if there is an edge connecting v and w; S_v and S_w divide the model into 2 clusters: $S_v = 1$ if v belongs to cluster 1, $S_v = -1$ if v belongs to cluster 2. The domain of Q belongs to [0, 1], in which the sum reaches 1 if modules clearly separate from each other, otherwise it reaches 0 if module division is faint. Identification of maximum network modularity is a NP-hard problem.

Therefore, the optimization algorithm has been enhanced to search the maximum modularity function of random graph partitions from the result of running 100 times of the algorithm in [49].

After the construction of the network, Cytospace was applied to extract the maximum interconnected network components, specifically 15,499 nodes and 64,944 edges (Figure 2). Finally, 49 modules (clusters) from the maximum connected network components were detected by the module optimization algorithm, where the clusters contained similar record and each cluster extracted each cancer rule. Based on the information from analyzing 49 clusters, the conclusions are provided in the next section.

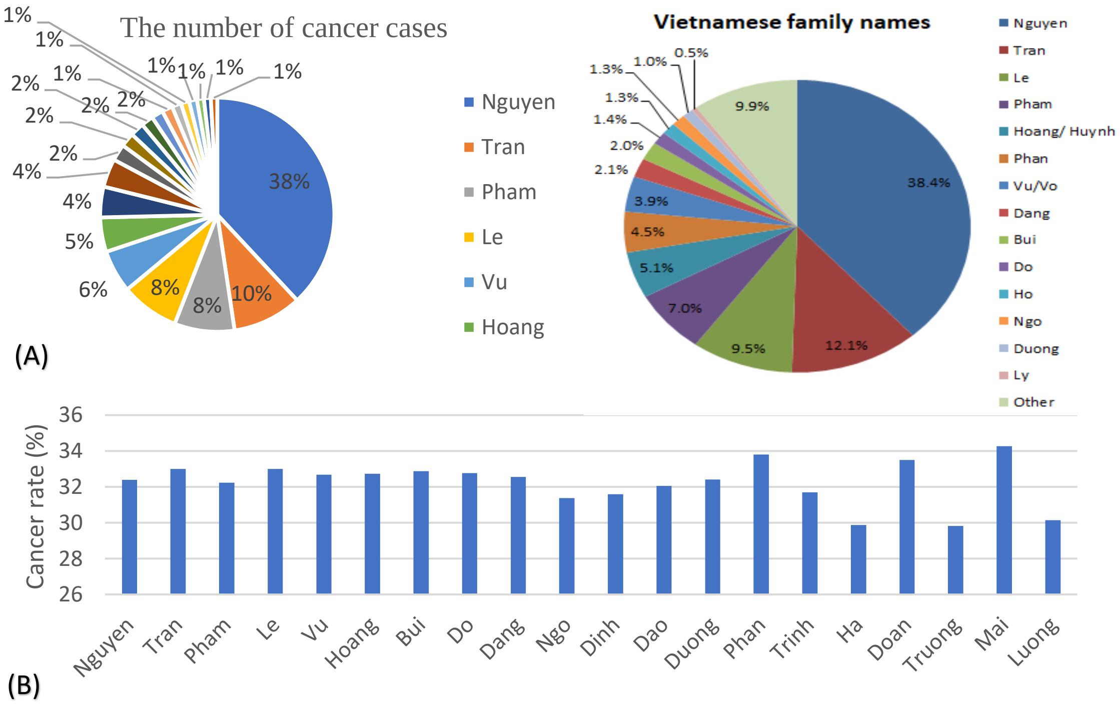

Figure 3. Cancer screening data tested by Family name. A) The number of cancer cases tested by Family name. The bar chart provides the cancer frequency by Family names while the Vietnamese population map in 2020 is provided by the pie chart. B) Cancer rate tested by Family name. There is not much difference between each family name in cancer rate with around 32 % of each, which Mai accounts for the highest rate (34.26 %) while Truong and Ha account for the lowest rate (29.88 %).

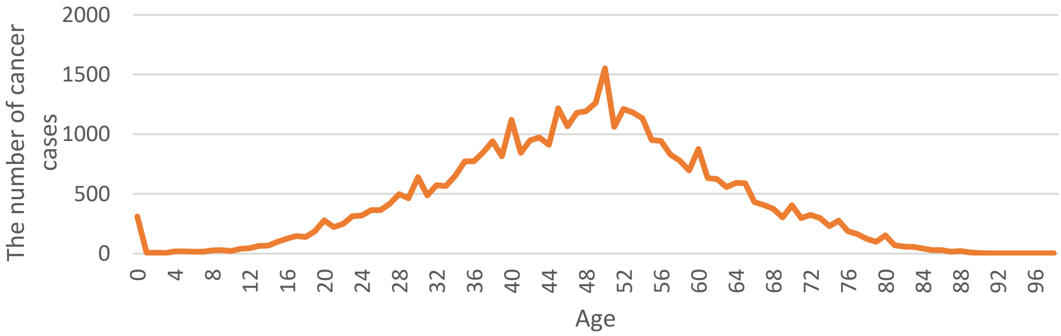

Figure 4. Cancer screening test data by Age.

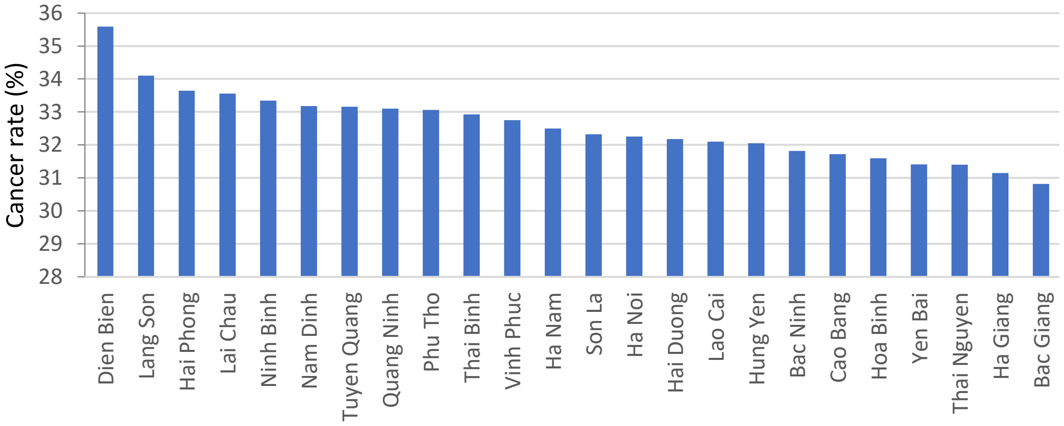

Figure 5. Cancer screening test data by Province.

3.4. Cancer rules

Firstly, the cancer screening test data was analyzed by the gender of people diagnosed with cancer. Most of the cancer rates were breast cancer in women compared to prostate cancer in men in Ha Noi. Regarding cancer rate, within 100 cases from each examination for both genders, it was 36 % for women and 34.6 % for men, respectively. Thus, women should take measures to improve their health to reduce the risk of cancer, such as eating healthier, consuming more vegetables and fruits, as well as maintaining a healthier lifestyle [50 - 52].

Secondly, we examined the data of cancer screening test of people diagnosed with cancer according to their family line (Figure 3). It is clear from Figure 3A that Nguyen, Tran, Pham and Le are the family lines of the Kinh ethnic group with the most prominent cancer rates, the second, third, and fourth highest corresponding to the proportion of the Vietnamese population map. As a result, there is a consistency between two different data sources, increasing the quality of the cancer screening test data. The Kinh is the largest ethnic group in Viet Nam, so the Kinh family lines also account for the majority (about 95 % of the population of Viet Nam), leading to a higher incidence and being affected by cancer and other diseases. In contrast, the family lines of ethnic minorities often live in mountainous areas, with a fresh climate and low population density, so they have very low rates of cancer. Figure 3B illustrates that although the number of cancer cases by family line is different when it comes to the family line, the number of cases is quite uniform, approximately 32 %. That shows that the probability of having cancer in the family is not significantly different. Previous research results have also shown that some cancers can be transmitted through genetic factors [11, 53 - 55].

Thirdly, Figure 4 illustrates the cancer screening test data by Age. We examined the age group between 0 - 99, in which the youngest age that can be affected was 0 and the oldest age was 99. Specifically, in the working-age under 60 years old, the age group 45 - 59 has the highest incidence. This is the oldest age group in the 3 working age groups, with weaker resistance and weaker immune system because they are older, so cancer cells have the ability to activate and spread quickly, creating favorable conditions for the development of cancer. Therefore, people in this age group need to adjust their eating habits in a healthy and nutritious way [52, 56, 57] to enhance the body's resistance as well as enhance individual immunity to cancer cells.

Fourthly, the cancer screening test data was analyzed by Address. After the investigation process, large provinces with large population and high population density are usually provinces with high cancer incidence, which are favorable for medical examination and treatment at the Vietnam Central Cancer Hospital. Dien Bien is the province with the highest cancer incidence in the country (about 35.59 %) although the cancer screening rate in this province is not high. The main cause may be due to over-exploitation of resources, leading to water pollution, increasing the risk of cancer (Figure 5). Provinces such as Ha Noi, Bac Ninh, Hung Yen, Hai Duong, Hai Phong, and Quang Ninh in the Red River Delta also have higher cancer rates than other regions. These provinces have many developing industrial zones, dense population, air pollution, water pollution, lack of trees, and unreasonable use of pesticides, all of which lead to a higher probability of human-borne infectious diseases compared to the upland provinces. Therefore, strictly controlling the discharges of industrial zones, limiting mineral exploitation, improving the quality of domestic air and water resources, and using pesticides properly may reduce cancer risk [20, 58 -62].

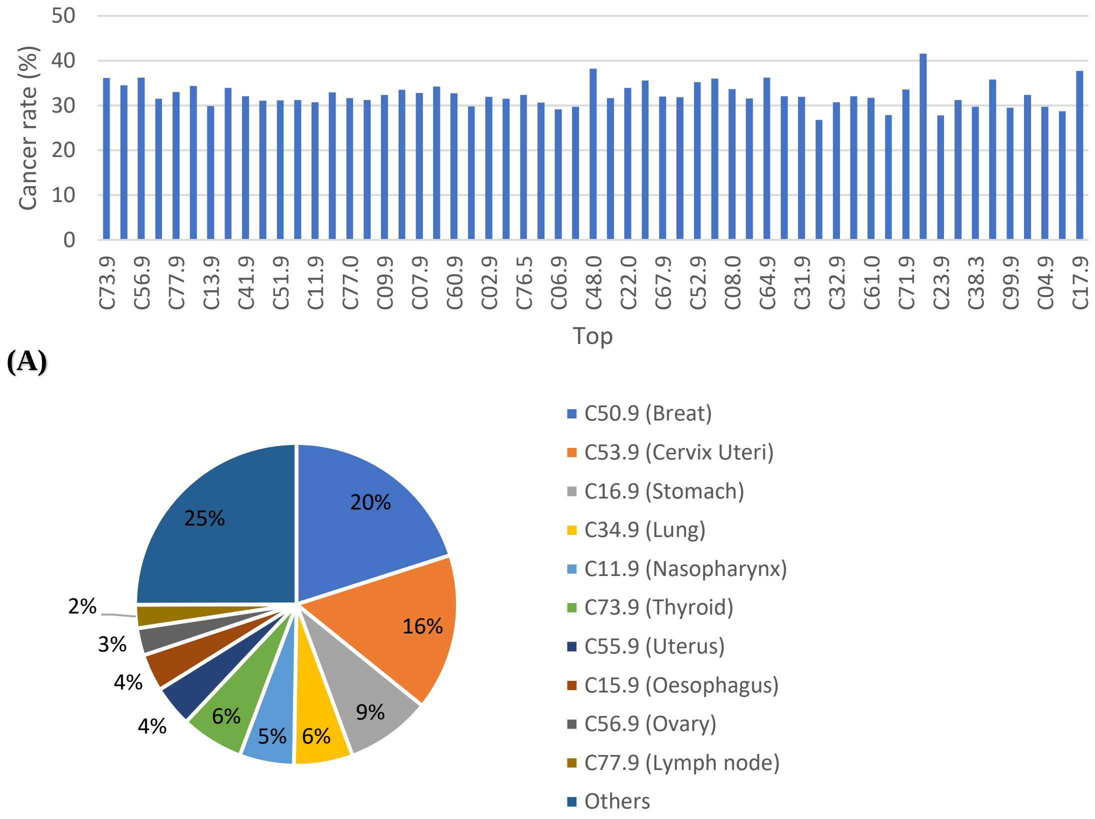

Figure 6. Cancer screening test data by type. Cancer type by percentage (**A**). 11 popular cancer types (**B**).

Fifthly, we examined cancer screening test data by Top (Figure 6). The most common cancer types were Breast cancer (C50.9) and cervix-uteri cancers (C53.9). The reason is due to women's hormonal changes as well as their complex reproductive structure, leading to the potential to produce cancer cells such as thyroid, breast, cervix, and ovary. Cancer groups C16.9 (stomach) and C34.9 (lung) were fairly common. The main reason is due to food safety and hygiene, widespread consumption of alcoholic beverages and tobacco, and lack of labor protections in a dangerous working environment. Therefore, it is important to raise people’s awareness of food safety and limit alcohol and tobacco consumption [63 - 67].

Finally, after examining the cancer screening test data, we drew the following conclusions:

1) Of the 100 cases tested by gender, 36 % and 34.6 % for women and men, respectively, were positive for cancer.

Women from 28 to 55 years old, starting with the surname Nguyen, Tran, Pham, Le, Vu, Bui, Hoang, and Do and living in the provinces of Ha Noi, Ninh Binh, Thai Binh, Hai Phong, Bac Ninh, Hai Duong, Hung Yen had a high-risk incidence for C53.9 (cervix uteri), C50.9 (breast), and C56.9 (ovarian).

Men between 50 and 61 years old, belonging to the Nguyen, Tran, and Pham families, living in Ha Noi and Thai Binh had a high-risk incidence for C13.9 (hypopharynx).

2) The larger the family lines, the higher the number of people getting cancer effect. In general, however, the rate of cancer correlating with family lines did not significantly differ among the 100 cases indicated for evaluation of each family line. The Mai family line had the highest rate (34.26 %), while the rates of the Truong and Ha families were the lowest (29.88%)

3) The majority of older people between the ages of 44 and 60 had the highest risk of cancer.

4) Provinces with high population density, such as Ha Noi, had a high number of cancer cases. Dien Bien province had the highest rate compared to others, about 33.5 - 36 %. The other provinces comprise 31 - 33 % out of the 100 people selected for the examination by province.

5) C37.9 (malignant renal) was the cancer with the highest rate (41.57%), followed by C48.0 (peritoneal) (38.23 %), and C17.9 (small intestine) (37.73 %). The lowest was C54.9 (malignant neoplasm of the corpus uteri) (26.8 %). Generally speaking, these figures fluctuated above 30 % of the 100 cancer screening tests for each type of cancer indicated (Figure 6A).

The common cancers in both men and women were the thyroid gland, lymph node, nasopharynx, esophagus, stomach, bronchus, and lung.

The incidence of cancers such as thyroid gland, lymph node, and nasopharynx was higher in women than in men. Meanwhile, the incidence of esophageal, stomach cancer, bronchus, and lung cancers was higher in men than in women.

6) After analyzing 190 cancers in the dataset, 75 % of the total number of recorded cancers belonged to 11 common cancer types, including breast, uterus, ovaries, liver, stomach, esophagus, colon, kidney, thyroid, oropharynx, bone, skin, lung, rectum, bladder, and prostate (Figure 6B).

4. CONCLUSIONS

A seven-attribute relation containing cancer screening test information from K-Hospital was exploited in this study. After the relation was pre-processed, we created a weighted undirected graph (network), where each node included a record, and each edge and weight represented the similarity between two data records. The discovery of novel cancer rules from the clusters for a deeper understanding of the present cancer circumstance in Viet Nam has been made after the application of complex network clustering to this weighted undirected graph using the network modularity optimization algorithm. It is disclosed from these cancer rules that some types of cancer run in certain family lines and living places in Viet Nam. The findings of the study can be used to support cancer diagnosis and screening, as well as to assist doctors and healthcare systems in making test orders based on patient data such as address, family, age group, and sex. Besides promoting the quality of medical care, environment, and nutrition, a serious and transparent public investment in data collection to build an advanced data capacity for analysis and prediction should be encouraged nationwide. It is shown from the study that immediate action should be taken with appropriate policies for the prevention as well as treatment of cancer patients. A comprehensive approach to tackling the problems caused by cancer, together with the engagement of different collaborators, could be more complicated than working within the healthcare system itself, but maybe the best choice for Viet Nam in the fight against the disease.

Acknowledgements. This research was funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2018.304.

CRediT authorship contribution statement. Minh Tan Nguyen (MTN): collected and processed the data, provided technical support for the study, drafted the manuscript. Duc Tinh Pham (DTP): analyzed the data and drafted the manuscript. Viet Ha Tran (VHT) technically supported the manuscript. Tien-Dzung Tran (TDT): conceived of the study, edited, reviewed, and gave guidance on the theoretical and mathematical issues and contexts. MTN and TDT contributed to critical feedback and editing.

Declaration of competing interest. We declare that we have no competing interests.

REFERENCES

1. Sung H., *et al*. - Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA: A Cancer Journal for Clinicians **71** (3) (2021) 209-249.
2. Tran T. D. and Pham D. T. - Identification of anticancer drug target genes using an outside competitive dynamics model on cancer signaling networks, Scientific Reports, **11** (1) (2021) 14095.
3. Thi Nguyen D. N., *et al.* - The burden of cervical cancer in Vietnam: Synthesis of the evidence, Cancer Epidemiology **59** (2019) 83-103.
4. Van Minh H., Van Thuan T., and Shu X. O. - Scientific Evidence for Cancer Control in Vietnam, Cancer Control **26** (1) (2019) 1073274819866450.
5. Pham T., *et al.* - Cancers in Vietnam - Burden and Control Efforts: A Narrative Scoping Review, Cancer Control **26** (1) (2019) 1073274819863802.
6. Nguyen S. M., *et al.* - Projecting Cancer Incidence for 2025 in the 2 Largest Populated Cities in Vietnam. Cancer Control **26** (1) (2019) 1073274819865274.
7. Cao B., *et al.* - Benchmarking life expectancy and cancer mortality: global comparison with cardiovascular disease 1981-2010, BMJ **357** (2017) j2765.
8. Mercurio V., *et al.* - Redox Imbalances in Ageing and Metabolic Alterations: Implications in Cancer and Cardiac Diseases. An Overview from the Working Group of Cardiotoxicity and Cardioprotection of the Italian Society of Cardiology (SIC), Antioxidants **9** (7) (2020) 641.
9. Tran T. D. and Kwon Y. K. - The relationship between modularity and robustness in signalling networks, J. R. Soc Interface **10** (88) (2013) 20130771.
10. Richiardi L., Pettersson A., and Akre O. - Genetic and environmental risk factors for testicular cancer, International Journal of Andrology **30** (4) (2007) 230-241.

11. BÁEz A. - Genetic and Environmental Factors in Head and Neck Cancer Genesis, Journal of Environmental Science and Health, Part C **26** (2) (2008) 174-200.
12. Ekman P. - Genetic and Environmental Factors in Prostate Cancer Genesis: Identifying High-Risk Cohorts, European Urology **35** (5-6) (1999) 362-369.
13. Goossens N., *et al.* - Cancer biomarker discovery and validation, Translational cancer research **4** (3) (2015) 256-269.
14. Tran T. D. and Kwon Y. K. - Hierarchical closeness efficiently predicts disease genes in a directed signaling network, Comput Biol. Chem. **53pb (**2014) 191-197.
15. Tran T. D. and Kwon Y. K. - Hierarchical closeness-based properties reveal cancer survivability and biomarker genes in molecular signaling networks, PLOS ONE **13** (6) (2018) e0199109.
16. Zeka A., Gore R., and Kriebel D. - Effects of alcohol and tobacco on aerodigestive cancer risks: a meta-regression analysis, Cancer Causes Control **14** (9) (2003) 897-906.
17. Castellsagué X., *et al.* - Independent and joint effects of tobacco smoking and alcohol drinking on the risk of esophageal cancer in men and women, Int J. Cancer **82** (5) (1999) 657-64.
18. Pöschl G. and Seitz H. K. - Alcohol and cancer, Alcohol and Alcoholism **39** (3) (2004) 155-165.
19. White A. J., *et al.* - Breast cancer and exposure to tobacco smoke during potential windows of susceptibility, Cancer Causes & Control **28** (7) (2017) 667-675.
20. Griffith J., *et al.* - Cancer Mortality in U.S. Counties with Hazardous Waste Sites and Ground Water Pollution, Archives of Environmental Health: An International Journal **44** (2) 91989) 69-74.
21. Morris R. D. - Drinking water and cancer. Environmental Health Perspectives **103** (suppl 8) 91995) 225-231.
22. Eichelberger L., *et al.* - Risk of Gastric Cancer by Water Source: Evidence from the Golestan Case-Control Study, Plos one **10** (5) 92015) e0128491.
23. Vanamala J. - Food systems approach to cancer prevention, Critical Reviews in Food Science and Nutrition **57** (12) 92017) 2573-2588.
24. Schwingshackl L., *et al.* - Food groups and risk of colorectal cancer, International Journal of Cancer **142** (9) (2018) 1748-1758.
25. Eckel S. P., *et al.* - Air pollution affects lung cancer survival, Thorax **71** (10) (2016) 891-898.
26. Turner M. C., *et al.* - Ambient Air Pollution and Cancer Mortality in the Cancer Prevention Study II, Environmental Health Perspectives **125** (8) (2017) 087013.
27. Wilding S., *et al.* - Decision regret in men living with and beyond nonmetastatic prostate cancer in the United Kingdom: A population-based patient-reported outcome study, Psycho-Oncology **29** (5) (2020) 886-893.
28. Kvåle K., Haugen D. F., and Synnes O. - Patients' illness narratives -From being healthy to living with incurable cancer: Encounters with doctors through the disease trajectory, Cancer Reports **3** (2) (2020) e1227.

29. Song P., Wu L., and Guan W. - Dietary Nitrates, Nitrites, and Nitrosamines Intake and the Risk of Gastric Cancer: A Meta-Analysis, Nutrients **7** (12) (2015) 9872-9895.
30. Joossens J. V., *et al.* - Dietary Salt, Nitrate and Stomach Cancer Mortality in 24 Countries, International Journal of Epidemiology **25** (3) (1996) 494-504.
31. Hertog M. G., *et al.* - Dietary flavonoids and cancer risk in the Zutphen Elderly Study, Nutr Cancer **22** (2) (1994) 175-84.
32. Wang M., *et al.* - A Review on Flavonoid Apigenin: Dietary Intake, ADME, Antimicrobial Effects, and Interactions with Human Gut Microbiota, BioMed Research International **2019** (2019) 7010467.
33. Mendonça L. A. B. M., *et al.* - The Complex Puzzle of Interactions Among Functional Food, Gut Microbiota, and Colorectal Cancer, Frontiers in Oncology **8 (**2018).
34. Scott L., Mobley L. R., and Il'yasova D. - Geospatial Analysis of Inflammatory Breast Cancer and Associated Community Characteristics in the United States, International Journal of Environmental Research and Public Health **14** (4) (2017) 404.
35. Truong C. D., Tran T. D., and Kwon Y. K. - MORO: a Cytoscape app for relationship analysis between modularity and robustness in large-scale biological networks, BMC Systems Biology **10** (4) (2016) 122.
36. Eide P. W., *et al.* - CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models, Scientific Reports **7** (1) (2017) 16618.
37. Jung Y. G., Kang M. S., and Heo J. - Clustering performance comparison using K-means and expectation maximization algorithms, Biotechnology & Biotechnological Equipment **28** (sup1) (2014) S44-S48.
38. Dubey A. K., Gupta U., and Jain S. - Analysis of k-means clustering approach on the breast cancer Wisconsin dataset, International Journal of Computer Assisted Radiology and Surgery **11** (11) (2016) 2033-2047.
39. Kakushadze Z. and Yu W. - *K-means and cluster models for cancer signatures, Biomolecular Detection and Quantification **13** (2017) 7-31.
40. Khan I., *et al.* - Ensemble clustering using extended fuzzy k-means for cancer data analysis, Expert Systems with Applications **172** (2021) 114622.
41. Sinaga K. P. and Yang M. S. - Unsupervised K-Means Clustering Algorithm, IEEE Access **8** (2020) 80716-80727.
42. Singh A., Yadav A., and Rana A. - K-means with three different distance metrics, International Journal of Computer Applications **67** (10) (2013).
43. Sneath P. H. A. - A method for testing the distinctness of clusters: A test of the disjunction of two clusters in Euclidean space as measured by their overlap, Journal of the International Association for Mathematical Geology **9** (2) (1977) 123-143.
44. Sneath P. H. A. - Basic program for a significance test for two clusters in euclidean space as measured by their overlap, Computers & Geosciences **5** (2) (1979) 143-155.
45. Sony A., *et al.* - Video summarization by clustering using euclidean distance, in 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies, 2011.
46. Hathaway R. J. and Bezdek J. C. - Nerf c-means: Non-Euclidean relational fuzzy clustering, Pattern Recognition **27** (3) (1994) 429-437.

47. Zhang Z., Kaiqi H., and Tieniu T. - Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes, in 18th International Conference on Pattern Recognition (ICPR'06), 2006.
48. Barber M. J. - Modularity and community detection in bipartite networks, Physical Review E **76** (6) (2007) 066102.
49. Guimerà R., Sales-Pardo M., and Amaral L. A. N. - Modularity from fluctuations in random graphs and complex networks, Physical Review E **70** (2) (2004) 025101.
50. Key T. J. - Fruit and vegetables and cancer risk, British Journal of Cancer **104** (1) (2011) 6-11.
51. Hurtado-Barroso S., *et al.* - Vegetable and Fruit Consumption and Prognosis Among Cancer Survivors: A Systematic Review and Meta-Analysis of Cohort Studies, Advances in Nutrition **11** (6) (2020) 1569-1582.
52. Byers T., *et al*. - American Cancer Society Guidelines on Nutrition and Physical Activity for Cancer Prevention: Reducing the Risk of Cancer with Healthy Food Choices and Physical Activity, CA: A Cancer Journal for Clinicians **52** (2) (2002) 92-119.
53. Lynch H. T., *et al.* - Hereditary Factors in Cancer: Study of Two Large Midwestern Kindreds, Archives of Internal Medicine **117** (2) (1966) 206-212.
54. Lynch H. T., *et al.* - Hereditary Factors in Gynecologic Cancer, The Oncologist **3** (5) (1998) 319-338.
55. Newman B., *et al.* - Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families, Proceedings of the National Academy of Sciences **85** (9) (1988) 3044-3048.
56. Doyle C., *et al.* - Nutrition and Physical Activity During and After Cancer Treatment: An American Cancer Society Guide for Informed Choices, CA: A Cancer Journal for Clinicians **56** (6) (2006) 323-353.
57. Nitenberg, G. and B. Raynard, Nutritional support of the cancer patient: issues and dilemmas. Critical Reviews in Oncology/Hematology, 2000. **34**(3): p. 137-168.
58. Ebenstein, A., The Consequences of Industrialization: Evidence from Water Pollution and Digestive Cancers in China. The Review of Economics and Statistics, 2012. **94**(1): p. 186-201.
59. Zhang X. L., *et al.* - Research and control of well water pollution in high esophageal cancer areas, World journal of gastroenterology **9** (6) (2003) 1187-1190.
60. Zhang X., *et al.* - Esophageal cancer spatial and correlation analyses: Water pollution, mortality rates, and safe buffer distances in China, Journal of Geographical Sciences **24** (1) (2014) 46-58.
61. Chunhabundit R. - Cadmium Exposure and Potential Health Risk from Foods in Contaminated Area, Thailand, Toxicological Research **32** (1) (2016) 65-72.
62. Boffetta P. - Human cancer from environmental pollutants: The epidemiological evidence. Mutation Research/Genetic Toxicology and Environmental Mutagenesis **608** (2) (2006) 157-162.
63. Wilde G. J. S. - Effects of mass media communications on health and safety habits: an overview of issues and evidence, Addiction **88** (7) (1993) 983-996.

64. Lee C. H., *et al.* - Independent and combined effects of alcohol intake, tobacco smoking and betel quid chewing on the risk of esophageal cancer in Taiwan, International Journal of Cancer **113** (3) (2005) 475-482.

65. de Graaf L., *et al.* - Live and let live: Residents' perspectives on alcohol and tobacco (mis)use in residential care facilities, International Journal of Older People Nursing **n/a**(n/a): p. e12508.

66. Salaspuro M. - Interactions of alcohol and tobacco in gastrointestinal cancer, Journal of Gastroenterology and Hepatology **27** (s2) (2012) 135-139.

67. Andre K., *et al.* - Role of alcohol and tobacco in the aetiology of head and neck cancer: A case-control study in the doubs region of France, European Journal of Cancer Part B: Oral Oncology **31** (5) (1995) 301-309.