به نام خدا

پدید آورنده:

سیده ندا ساداتی

استاد مربوطه:

دکتر باقری

موضوع:

تحلیل شبکه های اجتماعی در ابزار Gephi

هدف

هدف این پروژه تحلیل و آنالیز مجموعه های داده به کمک تحلیل شبکه های اجتماعی می باشد. مجموعه دادگان مورد استفاده در این پروژه از سایت Kaggle استخراج شده و به هدف طرح سوال و سناریو و تحلیل آنها به کار گرفته شده اند. در این پروژه از نرم افزار Gephi برای آنالیز دادگان استفاده شده است.

۱- مقدمه

در تحلیل شبکه های اجتماعی یک شبکه را به صورت مجموعه ای از گره ها و یالهایی که آنها را به هم متصل میکنند در نظر میگیرند. ساختار شبکه های اجتماعی معمولا به شکل گراف بوده و تحلیل آن پیچیده است. انواع مختلفی از شبکه های اجتماعی مانند شبکه های دوستی، علاقمندی و همکاری وجود دارند که بررسی و تحلیل این روابط، اساس کار Social Network Analysis را تشکیل میدهند. این کار با استفاده از نظریه های شبکه و گراف صورت میگیرد. در این پروژه قصد داریم با استفاده از تحلیل شبکه های اجتماعی، پاسخی برای سوالات مطرح شده در مورد داده خود بیابیم، همچنین با کمک نظریه های مطرح شده در این زمینه تحلیل و آنالیز آماری بر روی داده انجام دهیم و گراف به دست آمده از داده خود را تحلیل نماییم. از ابزارهای معمول برای تحلیل شبکه های اجتماعی میتوان به Gephi، Pajek و کتابخانه پایتون به نام Networkx اشاره نمود.

۲- مجموعه دادگان

در این پروژه از دو مجموعه داده World Health Statistics و Sales Data Sample استفاده شده است. هر یک از این دو مجموعه داده به فیلد خاصی اختصاص دارند. در ادامه هر یک از این دو مجموعه داده شرح داده میشوند.

۲-۱ World Health Statistics:

این مجموعه داده جدیدترین آمار بهداشتی جهان را پوشش میدهد (کشورهای به رسمیت شناخته شده توسط سازمان بهداشت جهانی) دسته بندی های مختلفی در این مجموعه داده وجود دارند شامل:

- امید به زندگی
- مرگ و میر مادران
- مرگ و میر نوزادان و کودکان
- بیماری های مسری
- بیماری های غیر مسری و سلامت روحی-روانی
- سوء مصرف مواد
- سلامت جنسی و باروری
- مرگ و میر ناشی از آلودگی محیط زیست
- سوء مصرف مواد الکلی
- تعداد نیروی کار در حوزه سلامت

برای این پروژه از ترکیب چندین دسته بندی استفاده شده است، برای مثال ترکیب دو دسته بندی نرخ مصرف الکل و نرخ خودکشی افراد با هم و ترکیب تعداد نیروی کار در حوزه های مختلف سلامت که شامل پزشک، داروساز، ماما و دندانپزشک می‌باشند. جدول ۱ توضیحی از جداول استفاده شده در این پروژه می‌باشد.

جدول ۱- شرح ویژگی مجموعه داده world health statistics

نام مجموعه دادگان	نام فایل/جدول	عنوان فیلد	نوع فیلد	شرح محتوای فیلد
World Health Statistics 2020	medicalDoctors	Location	string	country
	medicalDoctors	period	int	time of gathering data
	medicalDoctors	Indicator	string	Explanation of table
	medicalDoctors	First tooltip	float	Medical doctors (per 10,000)

country	string	Location	nursingAndMidwife	**World Health Statistics 2020**
time of gathering data	int	period	nursingAndMidwife	
Explanation of table	string	Indicator	nursingAndMidwife	
Nursing and midwifery personnel (per 10,000)	float	First tooltip	nursingAndMidwife	
country	string	Location	pharmacists	
time of gathering data	int	period	pharmacists	
Explanation of table	string	Indicator	pharmacists	
Pharmacists (per 10,000)	float	First tooltip	pharmacists	
country	string	Location	dentists	
time of gathering data	int	period	dentists	
Explanation of table	string	Indicator	dentists	
Dentists (per 10,000)	float	First tooltip	dentists	
country	string	Location	alcoholSubstanceAbuse	
time of gathering data	int	period	alcoholSubstanceAbuse	
Explanation of table	string	Indicator	alcoholSubstanceAbuse	
Sex	string	Dim1	alcoholSubstanceAbuse	
Total (recorded+unrecorded) alcohol per capita (15+) consumption	float	First tooltip	alcoholSubstanceAbuse	
country	string	Location	crudeSuicideRates	
time of gathering data	int	period	crudeSuicideRates	
Explanation of table	string	Indicator	crudeSuicideRates	
Sex	string	Dim1	crudeSuicideRates	
Crude suicide rates (per 100 000 population)	float	First tooltip	crudeSuicideRates	

۲-۲ Sales Data Sample:

این مجموعه داده شامل اطلاعات یک خرده فروشی وسایل نقلیه می‌باشد. جدول ۲ ویژگیهای این جدول را نشان میدهد.

جدول ۲- شرح ویژگی مجموعه داده Sales Data Sample

نام مجموعه دادگان	نام فایل/جدول	عنوان فیلد	نوع فیلد	شرح محتوای فیلد
Sales Data Sample	Sales Data Sample	ORDERNUMBER	integer	شماره سفارش
	Sales Data Sample	QUANTITYORDERED	integer	تعداد سفارشات
	Sales Data Sample	PRICEEACH	integer	قیمت هر کدام
	Sales Data Sample	ORDERLINENUMBER	integer	شماره خط سفارش
	Sales Data Sample	SALES	integer	فروش ها
	Sales Data Sample	ORDERDATE	date	تاریخ سفارش
	Sales Data Sample	STATUS	string	وضعیت ها
	Sales Data Sample	QTR_ID	id	-
	Sales Data Sample	MONTH_ID	id	ماه
	Sales Data Sample	YEAR_ID	id	سال
	Sales Data Sample	PRODUCTLINE	string	خط تولید
	Sales Data Sample	MSRP	integer	-
	Sales Data Sample	PRODUCTCODE	string	کد محصول
	Sales Data Sample	CUSTOMERNAME	string	نام مشتری
	Sales Data Sample	PHONE	string	تلفن
	Sales Data Sample	ADDRESSLINE1	string	آدرس خط۱
	Sales Data Sample	ADDRESSLINE2	string	آدرس خط ۲
	Sales Data Sample	CITY	string	شهر
	Sales Data Sample	STATE	string	حالت
	Sales Data Sample	POSTALCODE	string	کدپستی
	Sales Data Sample	COUNTRY	string	کشور
	Sales Data Sample	TERRITORY	string	ناحیه
	Sales Data Sample	CONTACTLASTNAME	string	نام معامله کننده
	Sales Data Sample	CONTACTFIRSTNAME	string	نام خانوادگی معامله کننده
	Sales Data Sample	DEALSIZE	string	اندازه معامله

۳- سوالهای مطرح شده

در مجموع ۳ سوال برای مجموعه داده ها طرح شده اند که دو سوال مربوط به داده World Health Statistics و یک سوال مربوط به داده Sales Data Sample می‌باشد. در ادامه به ۳ سوال مطرح شده اشاره میکنیم.

۳-۱ سوال اول مربوط به داده Sales Data Sample:

- متن سوال: یک نمایندگی فروش وسایل نقلیه (شامل انواع اتوموبیل، کشتی، هواپیما، موتورسیکلت و قطار)، ترکیبی از محصولات خود را در چه کشوری به فروش برساند که فروش بیشتری نسبت به دیگر محصولات داشته باشد؟
- مخاطبان/ذی نفعان سوال: بازاریاب نمایندگی فروش
- موضوع نیازمندیهای قابل رفع از مخاطبان/ذی نفعان سوال: تمرکز بر فروش ترکیبی از محصولات که مشتریان واقعی برای آنها وجود دارد.

۳-۲ سوال دوم مربوط به داده World Health Statistics:

- متن سوال: ترکیبی از نیروهای درمانی شامل دکتر، دندانپزشک، ماما و داروساز را در چه کشورهایی قرار دهیم تا مطمئن شویم افراد آن منطقه سرویسهای درمانی را دریافت میکنند؟
- مخاطبان/ذی نفعان سوال: سازمان بهداشت جهانی و مردم جهان
- موضوع نیازمندیهای قابل رفع از مخاطبان/ذی نفعان سوال: تمرکز بر کشورهایی که نیاز به خدمات درمانی دارند.

۳-۳ سوال سوم مربوط به داده World Health Statistics:

- متن سوال: مصرف الکل در یک کشور چه تاثیری در خودکشی افراد آن کشور دارد؟
- مخاطبان/ذی نفعان سوال: بازاریاب نمایندگی فروش
- موضوع نیازمندیهای قابل رفع از مخاطبان/ذی نفعان سوال: بررسی تاثیر الکل بر اختلالات روانی و خودکشی افراد.

در بخش بعدی به سناریوهای مطرح شده برای هر یک از این سوالات اشاره خواهیم نمود.

۴- سناریوهای طراحی شده

۴-۱ سناریو مربوط به سوال اول:

- عنوان سناریو: کشف مشتریان واقعی هر دسته از محصولات در کشورهای مختلف
- مفهوم و تعریف گره ها در سناریو: سفارش
- مفهوم و تعریف رنگ گره ها در سناریو: کشور سفارش دهنده محصولات
- مفهوم و تعریف اندازه گره ها در سناریو: حجم سفارش
- مفهوم و تعریف یال ها در سناریو: محصولات هم دسته

۴-۲ سناریو مربوط به سوال دوم:

- عنوان سناریو: کدام کشورها نیاز به گروه خاصی از خدمات درمانی دارند؟
- مفهوم و تعریف گره ها در سناریو: کشورها
- مفهوم و تعریف اندازه گره ها در سناریو: تعداد انواع خدمات درمانی مورد نیاز
- مفهوم و تعریف یال ها در سناریو: وجود تشابه در نیازمندی به نوع خاصی از خدمات درمانی

توضیحات: کشورهایی که نرخ نیروی درمانیشان کمتر از حد تعیین شده معمول است برای این سناریو در نظر گرفته شده اند.

۴-۳ سناریو مربوط به سوال سوم:

- عنوان سناریو: بررسی تاثیر مصرف الکل بر خودکشی افراد
- مفهوم و تعریف گره ها در سناریو: کشورها
- مفهوم و تعریف اندازه گره ها در سناریو: نرخ مصرف الکل
- مفهوم و تعریف یال ها در سناریو: وجود شباهت در نرخ خودکشی (اختلاف کمتر مساوی ۵)

توضیحات: این بررسی برای افراد بالای ۱۵ سال انجام میشود.

۵- پیش پردازش

در این مرحله میبایست داده ها را با توجه به نیازمندیهای هر سناریو تغییر دهیم.

۵-۱ پیش پردازش داده برای سناریوی اول:

سناریوی اول مربوط به مجموعه داده Sales Data Sample می‌باشد. در ابتدا ستونهای اضافی را پاک میکنیم، سپس در ستون STATUS تنها سطرهایی که سفارشهای آن فرستاده شده است (shipped) را در نظر میگیریم و از سفارش های لغو شده یا در حال اجرا صرف نظر میکنیم. سپس برای یکپارچگی بیشتر، تنها سفارشهای مربوط به جدیدترین سال یعنی ۲۰۰۵ را نگهداری میکنیم و باقی سطرها را حذف مینماییم. در پایان نیز شماره سفارشهای تکراری را حذف مینماییم. حال دیتاست را در فایلی به نام one.csv ذخیره میکنیم.

۵-۲ پیش پردازش داده برای سناریوی دوم:

سناریوی دوم مربوط به مجموعه داده World Health Statistics می‌باشد. در این سناریو میبایست ۴ جدول متفاوت را با هم ترکیب سازیم. برای این منظور ابتدا تک تک جداول را یکپارچه ساخته و در ادامه آنها را merge مینماییم. جدول اول مربوط به تعداد دندانپزشکان در هر ۱۰ هزار نفر جمعیت می‌باشد. در ابتدا ستون Indicator که دارای یک مقدار واحد می‌باشد و توضیحی از جدول ارائه میکند را حذف مینماییم. سپس سطرهای مربوط به سالهای ۲۰۱۷ و ۲۰۱۸ را برای یکسان سازی داده نگهداری کرده و باقی سطرها را حذف میکنیم. در ادامه برای مشخص شدن کشورهای نیازمند به دندانپزشک، سطرهایی که حاوی تعداد دندانپزشک بیشتر از ۴.۹۸ می‌باشند را حذف میکنیم. در ادامه نام ستون First Tooltip که مربوط به تعداد دندانپزشکان می‌باشد را به dentist تغییر میدهیم تا در ادامه با ترکیب جداول، ستونها قابل تفکیک باشند.

سپس تمام این مراحل را برای جداول پزشک، ماما و داروساز انجام میدهیم و برای هر یک کشورهای نیازمند به نیروی کار را با محدود کردن تعداد سطرها مشخص مینماییم. در انتها تمام جداول بدست آمده را با دستور pd.merge(df1, df2, how='outer') یکسان میکنیم و جدول نهایی را در فایلی با نام two.csv ذخیره مینماییم.

۵-۳ پیش پردازش داده برای سناریوی سوم:

سناریوی سوم نیز مربوط به مجموعه داده World Health Statistics می‌باشد. در این سناریو میبایست ۲ جدول متفاوت را با هم ترکیب سازیم. برای این منظور ابتدا هر جدول را یکپارچه ساخته و سپس آنها را merge مینماییم. ابتدا جدول مربوط به نرخ خودکشی افراد را خوانده و ستون Indicator که دارای یک مقدار

واحد می‌باشد و توضیحی از جدول ارائه میکند را حذف مینماییم. سپس سطرهای مربوط به سالهای به غیر از ۲۰۱۵ را حذف مینماییم. در ادامه در ستون Dim1 که به جنسیت افراد اشاره دارد، تنها سطرهایی که مربوط به هر دو جنسیت می‌باشد را نگه داشته و باقی سطرها را حذف میکنیم. سپس نام ستون First Tooltip که مربوط به نرخ خودکشی افراد می‌باشد را به crudeSuicideRate تغییر میدهیم تا در ادامه با ترکیب جداول، ستونها قابل تفکیک باشند. در ادامه ستون Dim1 را حذف میکنیم زیرا حاوی اطلاعات مفیدی برای ادامه پروژه نیست. تمام این مراحل را برای جدول مربوط به نرخ سوء مصرف الکل انجام میدهیم و در پایان این دو جدول را با دستور pd.merge(df1, df2, how='inner') ترکیب میکنیم و جدول نهایی را در فایلی با نام three.csv ذخیره میکنیم.

پس از انجام پیش پردازش داده ها، میبایست آنها را به فرمتی تبدیل کنیم تا آماده ورود به ابزار gephi باشند، برای این منظور برای هر سناریو یک فایل گره به نام Node و یک فایل یال به نام Edge ایجاد کرده و جزئیات گره ها و یالها را مانند سایز و رنگ اعمال میکنیم. جدول Node دارای ستونهایی با نام Id، Label، Color و Size بوده و جدول Edge دارای ستونهایی با نام Source، Target، Type و Weight می‌باشد.

در سناریوی اول label گره را با شماره سفارش مشخص کرده و رنگ گره را بر اساس کشور سفارش دهنده محصول و اندازه گره را با توجه به حجم سفارش تعیین میکنیم. در سناریوی دوم label گره نشان دهنده نام کشور و اندازه گره با توجه به تعداد انواع خدمات درمانی مورد نیاز آن کشور انتخاب میشود. در سناریوی سوم نیز label گره را با کشور و اندازه گره را با توجه نرخ مصرف الکل آن کشور تعیین میکنیم.

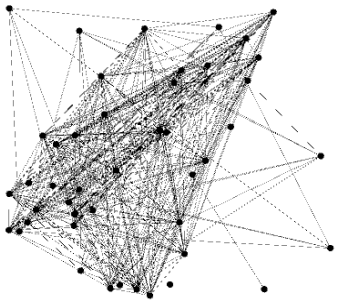
در نهایت برای سناریوی اول ۴۵ گره و ۳۷۰ یال، برای سناریوی دوم ۱۶۵ گره و ۱۲۷۰۵ یال و برای سناریوی سوم ۱۸۲ گره و ۸۰۹۴ یال ایجاد شد. یالهای تمامی سناریوها بدون جهت و با وزن ۱ بوده اند.

۶- روش انجام

در این مرحله فایلهای گره و یال را در ابزار gephi وارد کرده و شبکه به دست آمده را تحلیل مینماییم. در ادامه جزئیات پیاده سازی هر کدام از سناریوها آورده شده است.

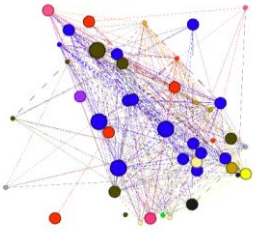
۱-۶ پیاده سازی سناریوی اول:

در سناریوی اول میبایست در مجموعه داده sales data sample مشتریان واقعی هر دسته از محصولات را در کشورهای مختلف کشف کنیم. پس از باز کردن برنامه gephi یک workspace جدید ایجاد میکنیم. سپس از قسمت Data Laboratory فایل گره ها و یالها را import میکنیم. شکل ۱ نمای اولیه شبکه را نشان میدهد.



شکل ۱- نمای اولیه شبکه

در ادامه از منوی Appearance در قسمت Partition رنگ گره ها را با توجه به مقادیر از پیش تعیین شده مشخص میکنیم. سپس سایز گره ها را از قسمت Size و بخش Ranking تغییر میدهیم. سایز گره ها با توجه به حجم سفارش از ۱۰ تا ۳۵ متغیر است. رنگ گره ها در جدول ۳ آورده شده اند. شکل ۲ شبکه را پس از اعمال تغییرات نشان میدهد.



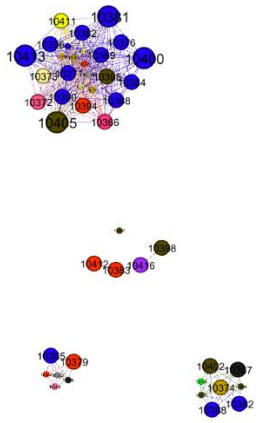
شکل ۲- نمای شبکه پس از اعمال تغییرات اولیه

جدول ۳- رنگ گره ها در سناریوی اول

نام کشور	رنگ گره
سنگاپور	سیاه
آمریکا	آبی
فرانسه	قهوه ای

فنلاند	کرمی
استرالیا	طلایی
اتریش	خاکستری
انگلستان	سبز
سوئد	نارنجی
ژاپن	صورتی
بلژیک	آلبالویی
ایتالیا	بنفش
اسپانیا	قرمز
کانادا	زرد

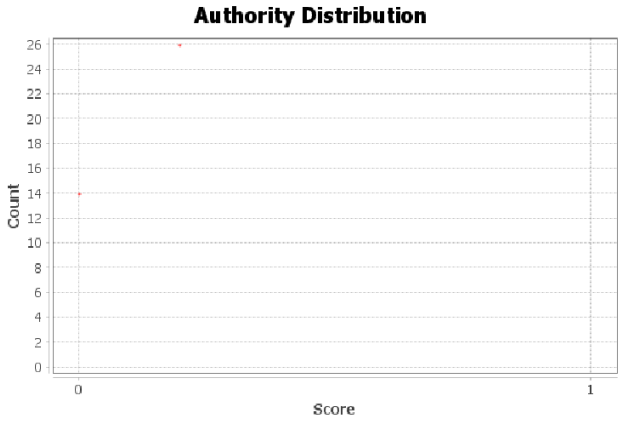
با استفاده از الگوریتمهای نمایش گراف میتوان شکل بهتری از شبکه را مشاهده کرد، برای این منظور از منوی Layout استفاده کرده و لی اوت Force Atlas را انتخاب میکنیم. سپس از لی اوت No overlap استفاده میکنیم تا شبکه را بدون درهم رفتگی گره ها مشاهده کنیم. در ادامه نام هر گره را که در جدول Nodes با نام Label تعیین کرده بودیم به شبکه اضافه میکنیم. شکل ۳ نمای شبکه را پس از اعمال لی اوت ها نشان میدهد.



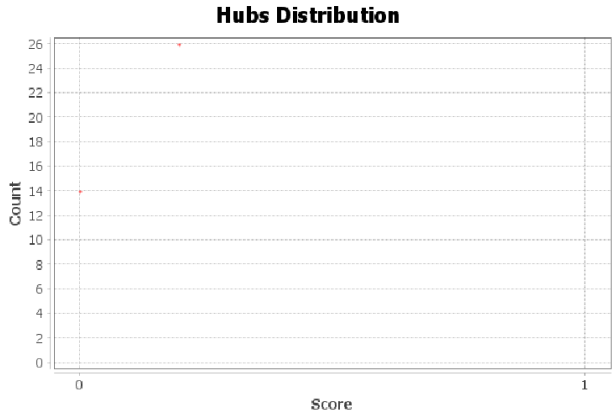
شکل ۳- نمای شبکه پس از اعمال لی اوت های force atlas و no overlap

در این تصویر یک گره بدون یال مشاهده میشود که مربوط به سفارش کشتی می‌باشد، همچنین سایز گره کوچک است که نشان دهنده حجم کم سفارش می‌باشد، میتوان نتیجه گیری کرد که کشتی محصول پرفروشی در میان دیگر محصولات نیست و بهتر است بر فروش آن تمرکز نکنیم. همچنین گره هایی با یک یال میان آنها مشاهده میشود، یکی از این دو گره مربوط به سفارش اتوبوس و کامیون می‌باشد، دو گره دیگر مربوط به سفارش هواپیما می‌باشد. حجم سفارش اتوبوس و کامیون و هواپیما متوسط می‌باشد که به این معنیست که نباید به طور کامل

خرید و فروش آن را کنار نهاد، بلکه باید توجه کمتری نسبت به سایر دسته بندیها به آن نشان دهیم اما همچنان آنها را در میان وسایل نقلیه با فروش متوسط قرار دهیم. در ادامه از مسیر Filters\Topology\DegreeRange گره های با یالهای کمتر از ۳ را فیلتر میکنیم تا شبکه ای خلوتتر به دست آید. در ادامه با کمک منوی Statistics اطلاعات آماری مربوط به شبکه را محاسبه میکنیم. مقدار average degree برابر ۱۸.۴ می‌باشد که به این معنیست که به طور متوسط هر گره ۱۸ یال دارد. قطر شبکه نیز ۱ به دست آمده که نشان میدهد طول بلندترین، کوتاه ترین مسیر در گراف ۱ می‌باشد. معیار بعدی برای کسب اطلاعات در مورد شبکه HITS می‌باشد که یک الگوریتم آنالیز لینک می‌باشد. در این الگوریتم دو مقدار برای هر گره محاسبه میشود، authority و hub. مقدار اول میزان ارزشمندی اطلاعات نهفته در یک گره و مقدار دوم میزان کیفیت لینکهای گره ها می‌باشد. شکل ۴ و ۵ مقدار به دست آمده برای authority و hub را نشان میدهد.



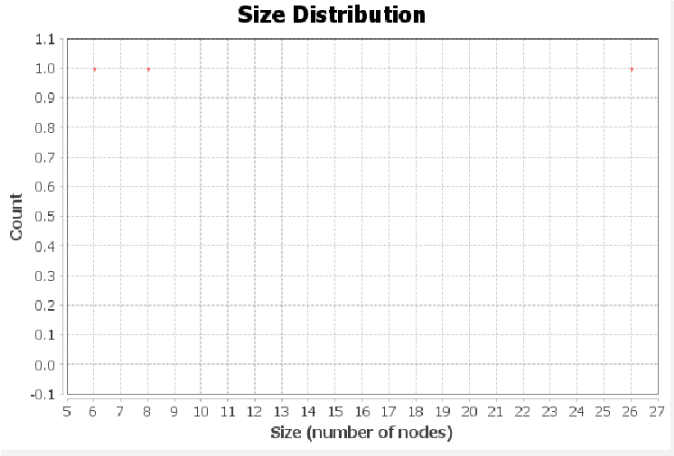
شکل ۴- نمودار authority



شکل ۵- نمودار hub

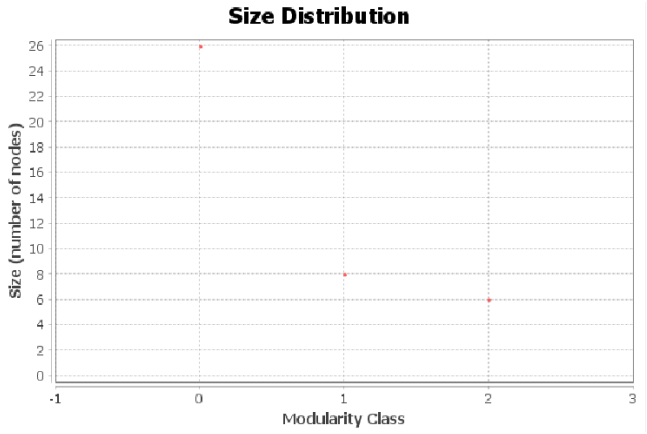
با توجه به دو نمودار به دست آمده میتوان دریافت که ۱۴ گره در شبکه، امتیاز و ارزشمندی برابر با ۰ دارند و ۲۶ گره ارزشمندی کمتر از ۰.۵ دارند. پس نتیجه میشود ۶۵ درصد گره ها و حدود ۴۰ درصد لینکهای شبکه ارزشمند هستند.

تعداد Weakly Connected Components های شبکه ۳ عدد می‌باشد. شکل ۶ توزیع گره ها در هر یک از کامپوننتها را نشان میدهد. طبق شکل میتوان دریافت که ۲۶، ۸ و ۶ گره در هر کامپوننت وجود دارند.



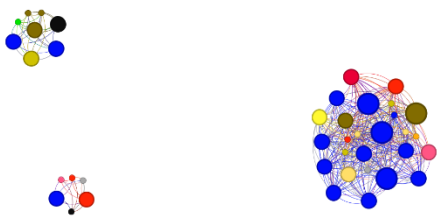
شکل ۶- توزیع گره ها در هر کامپوننت

طبق الگوریتم های تشخیص جوامع مانند modularity میتوان دریافت که ۳ جامعه (community) در گراف وجود دارد. شکل ۷ توزیع گره ها در هر جامعه را نشان میدهد که به ترتیب ۲۶، ۸ و ۶ گره در هر جامعه قرار دارند.



شکل ۷- توزیع گره ها در هر community

طبق اطلاعات به دست آمده از شکلهای ۶ و ۷ و همچنین گراف، میتوان دریافت شبکه دارای یک جامعه بزرگ بوده که مربوط به اتوموبیلهای کلاسیک (با قدمت ۲۰ ساله) می‌باشد. همچنین دو جامعه کوچکتر مربوط به موتورسیکلتها و اتوموبیلهای وینتج (قدیمی و مربوط به دهه های گذشته) می‌باشد. پس میتوان نتیجه گرفت خرید و فروش این ۳ دسته از وسایل نقلیه سود بیشتری به دنبال دارد. در نهایت از پنجره preview نمای نهایی گراف را به دست می آوریم. شکل ۸ preview گراف را نشان میدهد.

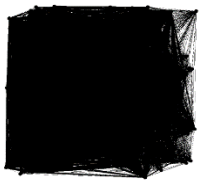


شکل ۸- preview گراف

۲-۶ پیاده سازی سناریوی دوم:

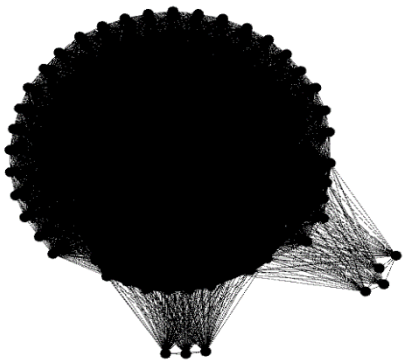
در سناریوی دوم باید به این سوال پاسخ دهیم که در مجموعه داده World Health Statistics کدام کشورها نیاز به گروه خاصی از خدمات درمانی دارند؟

در ابتدا یک workspace جدید ایجاد میکنیم. سپس از قسمت Data Laboratory فایل گره ها و یالها را import میکنیم. شکل ۹ نمای اولیه شبکه را نشان میدهد.



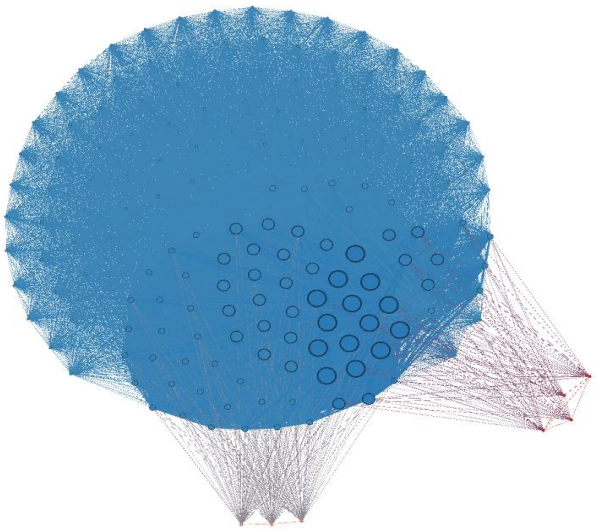
شکل ۹- نمای اولیه شبکه

در ادامه از منوی Appearance در قسمت Size و بخش Ranking اندازه گره ها را با توجه به تعداد نیازمندی هر کشور به نوع خاصی از نیروی کار درمان تغییر میدهیم و این مقادیر از ۵ تا ۳۰ متغیر است. سپس لی اوت force atlas و سپس no overlap را اعمال میکنیم. شکل ۱۰ نمای گراف را در این مرحله نمایش میدهد.



شکل ۱۰- نمای گراف پس از اعمال لی اوت

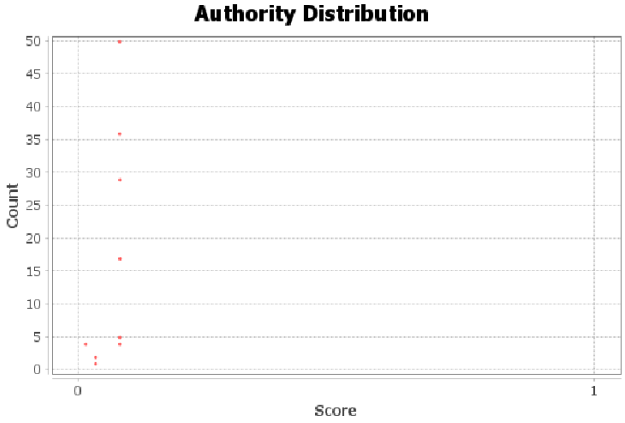
در ادامه برای بالا بردن خوانایی شبکه از منوی statistics گزینه average degree را انتخاب میکنیم تا میانگین درجات گراف را محاسبه کند. سپس میتوانیم از منوی appearance رنگ گره ها را با توجه به درجه شان تغییر دهیم. شکل ۱۱ نمای گراف را پس از اعمال تغییرات نشان میدهد.



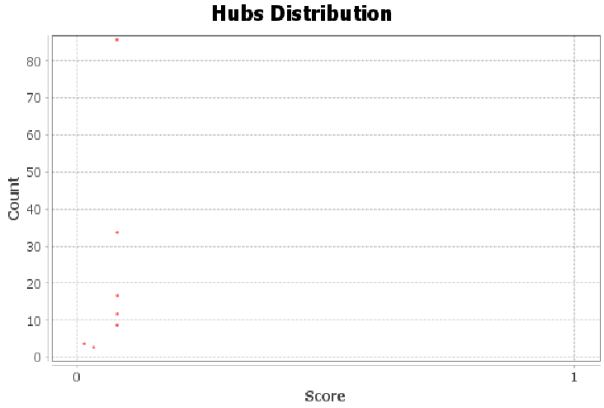
شکل ۱۱- نمای گراف پس از تغییر رنگ گره ها

در ادامه با کمک منوی Statistics اطلاعات آماری مربوط به شبکه را محاسبه میکنیم. مقدار average degree برابر ۱۵۴ می‌باشد که به این معنیست که به طور متوسط هر گره ۱۵۴ یال دارد. قطر شبکه نیز ۲ به دست آمده

که نشان میدهد طول بلندترین، کوتاه ترین مسیر در گراف ۲ می‌باشد. در ادامه با الگوریتم HITS مقادیر authority و hub را به دست می آوریم. مقدار اول میزان ارزشمندی اطلاعات نهفته در یک گره و مقدار دوم میزان کیفیت لینکهای گره ها می‌باشد. شکل ۱۲ و ۱۳ مقدار به دست آمده برای authority و hub را نشان میدهد. شکل ۱۲ نشان میدهد تعداد بسیار زیادی از گره ها ارزش اطلاعاتی بسیار پایین و نزدیک به صفر دارند. همچنین شکل ۱۳ نشان میدهد لینکهای بین گره ها نیز از کیفیت نزدیک به صفر برخوردارند. این به این معنیست که حدود ۲۰ درصد از گره ها و لینکهای گراف ما حاوی اطلاعات مفید برای آنالیز و تحلیل می‌باشند.

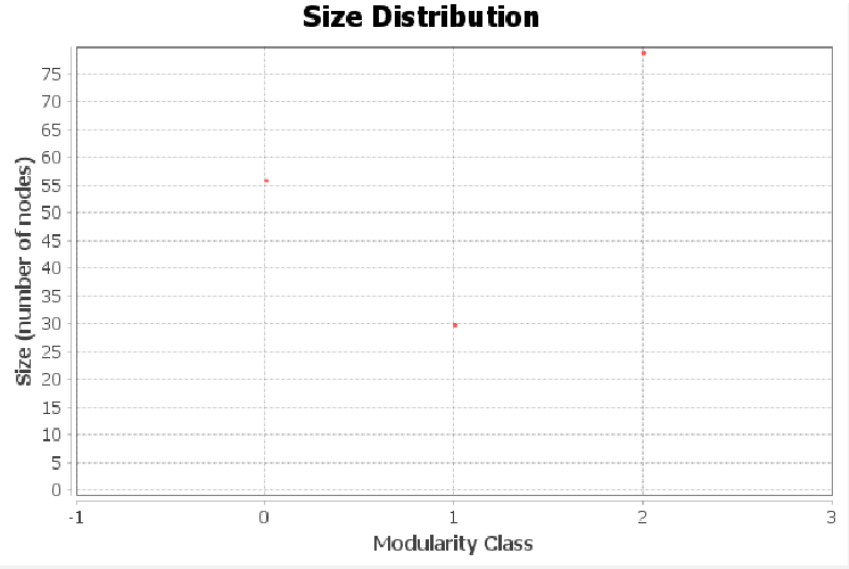


شکل ۱۲- نمودار authority



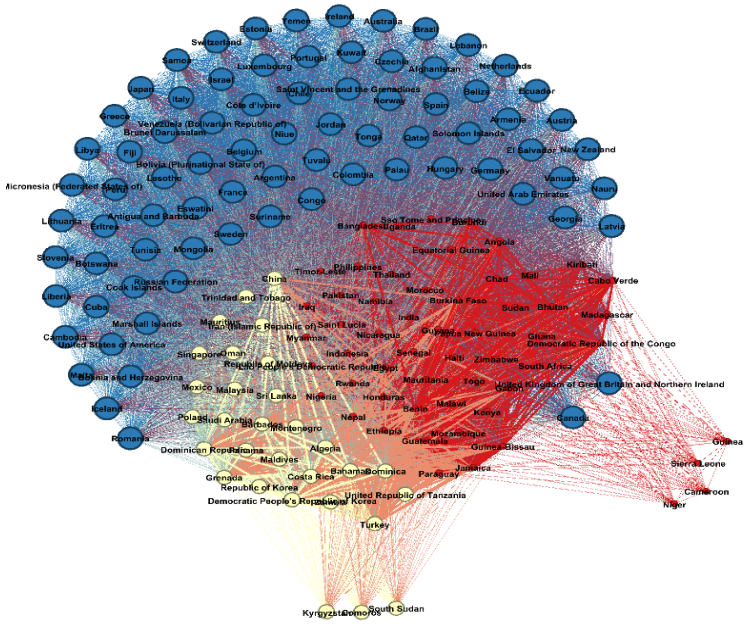
شکل ۱۳- نمودار hub

طبق الگوریتم های تشخیص جوامع مانند modularity میتوان دریافت که ۳ جامعه (community) در گراف وجود دارد. شکل ۱۴ توزیع گره ها در هر جامعه را نشان میدهد که به ترتیب ۸۰، ۵۵ و ۳۰ گره در هر جامعه قرار دارند.



شکل ۱۴- نمودار community های گراف

در ادامه از منوی appearance رنگ گره ها را بر اساس ranking و با توجه به خصیصه modularity class تغییر میدهیم. سپس اندازه گره ها را نیز با توجه به همین خصیصه یعنی modularity class تغییر میدهیم. حال گرافی با ۳ کلاس (community) داریم. هر کلاس با یک رنگ و اندازه مشخص شده است و خوانایی گراف بالاتر رفته است. شکل ۱۵ نمایی از گراف را نشان میدهد.



شکل ۱۵- تغییر رنگ و اندازه گره ها با توجه به جوامع موجود در گراف

از شکل میتوان دریافت که ۳ جامعه وجود دارد. اسامی برخی کشورهای درون هر جامعه در جدول ۴ قابل مشاهده است.

جدول ۴- کشورهای موجود در جوامع

رنگ گره های جامعه	نام کشورهای جامعه
قرمز	بنگلادش- فیلیپین- پاکستان- عراق- مصر- میانمار- آفریقای جنوبی- آنگولا- اندونزی- سنگال
کرمی	ایران- ترکیه- کره جنوبی- کره شمالی- مالزی- عمان- سنگاپور- چین- مکزیک- عربستان- سودان جنوبی
آبی	کانادا- انگلستان- سوئد- آلمان- امارات- روسیه- ایرلند- لیبی- ژاپن- ایتالیا- اسرائیل- یونان

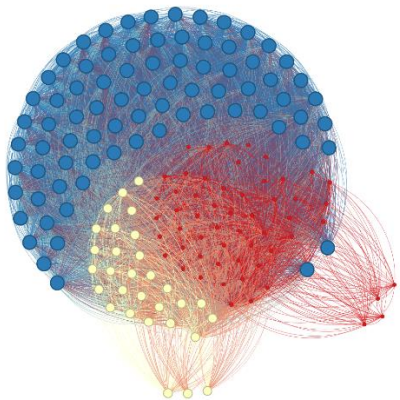
میزان نیازمندی هر جامعه به نیروی کادر درمان شامل ماما، پزشک، داروساز و دندانپزشک به صورت زیر است:

جامعه قرمز رنگ بین ۲ تا ۴ نیازمندی دارد، عموم این کشورها نیاز به ماما، دندانپزشک و پزشک دارند.

جامعه کرمی رنگ ۲ نیازمندی دارد که شامل دندانپزشک و ماما می‌باشند.

جامعه آبی رنگ فقط یک نیازمندی دارد و آن هم نیاز به ماما می‌باشد.

در پایان با استفاده از پنجره preview نمای نهایی گراف را نمایش میدهیم. در شکل ۱۶ preview گراف قابل مشاهده می‌باشد.

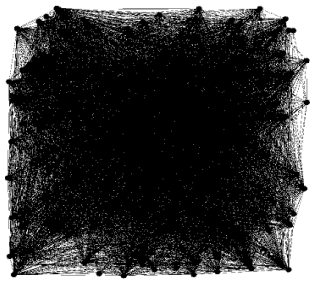


شکل ۱۶- preview گراف

۳-۶ پیاده سازی سناریوی سوم:

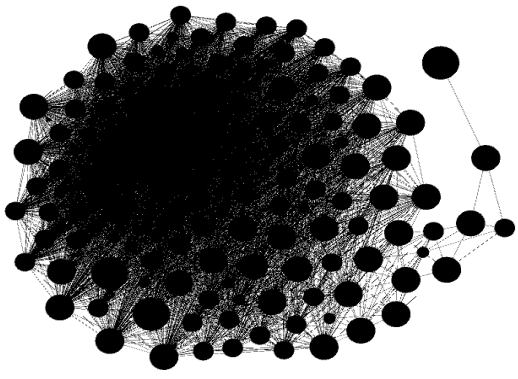
در سناریوی سوم باید به این سوال پاسخ دهیم که مصرف الکل در یک کشور چه تاثیری در خودکشی افراد آن کشور دارد؟

در ابتدا یک workspace جدید ایجاد میکنیم. سپس از قسمت Data Laboratory فایل گره ها و یالها را import میکنیم. شکل ۱۷ نمای اولیه شبکه را نشان میدهد.



شکل ۱۷- نمای اولیه شبکه

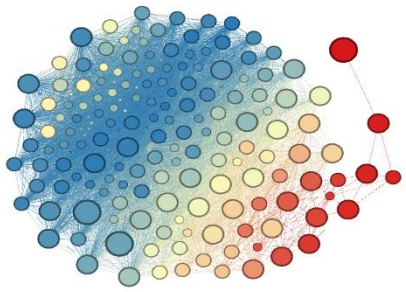
در ادامه از منوی Appearance در قسمت Size و بخش Ranking اندازه گره ها را با توجه به نرخ مصرف الکل هر کشور تغییر میدهیم و این مقادیر از ۵ تا ۴۵ متغیر است. سپس لی اوت yifan hu و سپس no overlap را اعمال میکنیم. شکل ۱۸ نمای گراف را در این مرحله نمایش میدهد.



شکل ۱۸- نمای گراف پس از اعمال لی اوت ها

در ادامه برای خواناتر شدن شبکه، مقدار average degree را محاسبه میکنیم. سپس از منوی appearance رنگ گره ها را با توجه به درجه آن تغییر میدهیم. گره های با درجه کمتر به رنگ قرمز و گره های با درجه بیشتر به رنگ آبی درآمده اند. از این شکل میتوان دریافت کشورهایی که نرخ مصرف الکل بالاتری دارند لزوما نرخ خودکشی

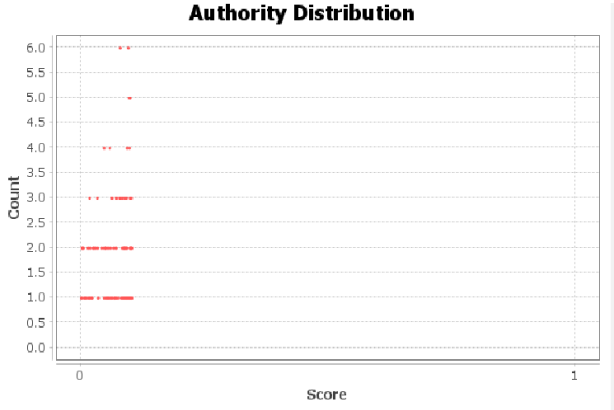
مشابهی ندارند و شباهت در نرخ خودکشی کشورها میتواند مربوط به دو کشور با نرخ مصرف الکل متفاوت باشد. شکل ۱۹ گراف را در این مرحله نشان میدهد.



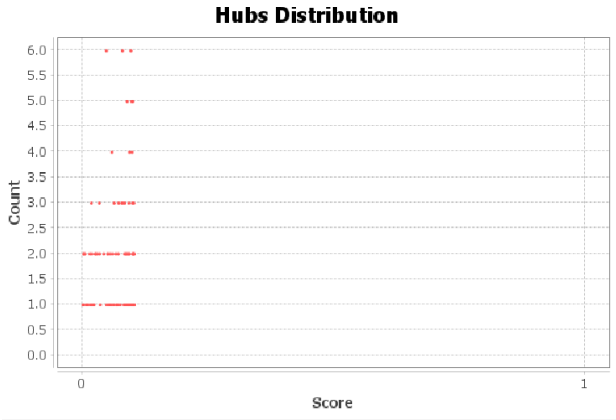
شکل ۱۹- تغییر رنگ گره ها با توجه به درجه آنها

در ادامه از طریق منوی filters گره های با درجه کمترمساوی ۵ را از گراف حذف میکنیم و اطلاعات آماری را برای گراف به دست می آوریم.

مقدار average degree برابر ۹۰.۳۶۹ می‌باشد که به این معنیست که به طور متوسط هر گره ۹۰ یال دارد. قطر شبکه نیز ۶ به دست آمده که نشان میدهد طول بلندترین، کوتاه ترین مسیر در گراف ۶ می‌باشد. در ادامه با الگوریتم HITS مقادیر authority و hub را به دست می آوریم. مقدار اول میزان ارزشمندی اطلاعات نهفته در یک گره و مقدار دوم میزان کیفیت لینکهای گره ها می‌باشد. شکل ۲۰ و ۲۱ مقدار به دست آمده برای authority و hub را نشان میدهد. شکل ۲۰ نشان میدهد تعداد بسیار زیادی از گره ها ارزش اطلاعاتی پایین و نزدیک به صفر دارند. درواقع گره ها حدودا ۲۰ درصد مقادیر ارزشمند دارند. شکل ۲۱ نشان میدهد لینکهای بین گره ها نیز از کیفیت نزدیک به صفر برخوردارند. این به این معنیست که حدود ۲۰ درصد از گره ها و لینکهای گراف ما حاوی اطلاعات مفید برای آنالیز و تحلیل می‌باشند.

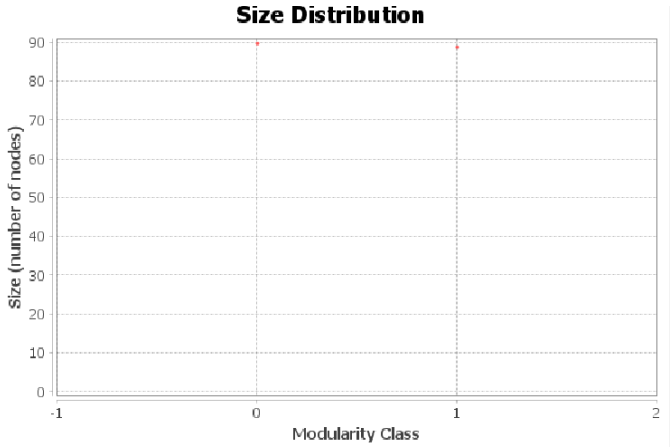


شکل ۲۰- نمودار authority گراف



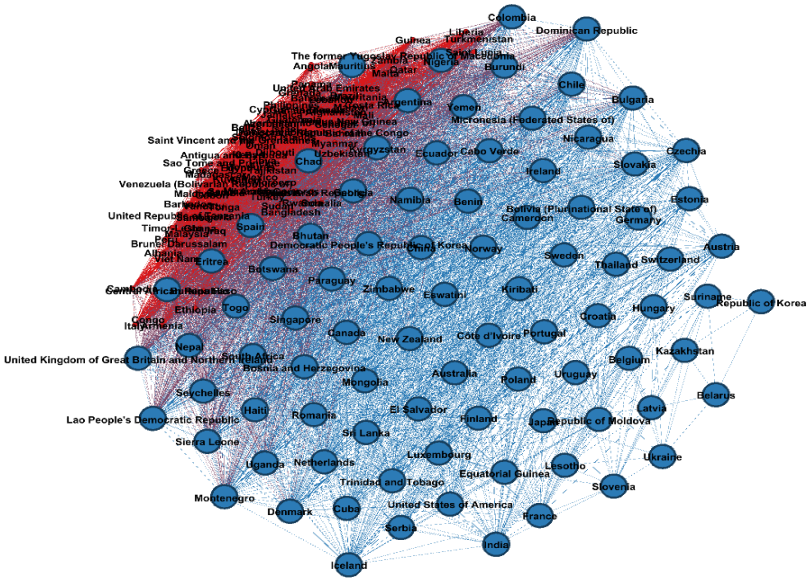
شکل ۲۱- نمودار hub گراف

طبق الگوریتم های تشخیص جوامع مانند modularity میتوان دریافت که ۲ جامعه (community) در گراف وجود دارد. شکل ۲۲ توزیع گره ها در هر جامعه را نشان میدهد که ۹۰ گره در هر جامعه قرار دارد.



شکل ۲۲- توزیع گره ها در community

در ادامه از منوی appearance رنگ گره ها را بر اساس ranking و با توجه به خصیصه modularity class تغییر میدهیم. سپس اندازه گره ها را نیز با توجه به همین خصیصه یعنی modularity class تغییر میدهیم. حال گرافی با ۲ کلاس (community) داریم و هر کلاس با یک رنگ و اندازه مشخص شده است. سپس با لی اوت no overlap خوانایی گراف را بالاتر برده ایم. شکل ۲۳ نمایی از گراف را نشان میدهد.



شکل ۲۳- تغییر رنگ و اندازه گره ها با توجه به جوامع موجود در گراف

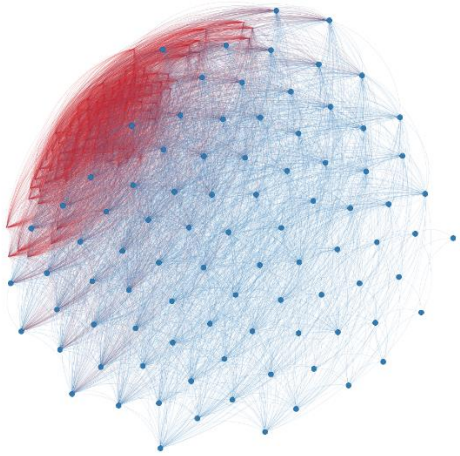
همانطور که پیشتر در قسمت اطلاعات آماری دیدیم گراف به ۲ جامعه با گره های یکسان تقسیم شده است. کلاس ۱ با رنگ آبی و کلاس ۲ با رنگ قرمز مشخص شده است. جدول ۵ به برخی کشورهای درون هر جامعه اشاره میکند.

جدول ۵- کشورهای درون هر جامعه

رنگ گره های جامعه	نام کشورهای جامعه
قرمز	ایران، عراق، مالزی، ونزوئلا، کویت، ترکیه، امارات، سومالی، میانمار، عمان
آبی	آرژانتین، اسلواکی، استرالیا، سوئیس، سوئد، آفریقای جنوبی، انگلستان، آلمان، کره جنوبی

پس از میانگین گیری از نرخ خودکشی و نرخ مصرف الکل در تمام کشورهای مورد بررسی، مشخص شد کشورهای جامعه قرمز رنگ عموما نرخ خودکشی و مصرف الکل پایینتر از میانگین جهان را دارند. همچنین کشورهای جامعه آبی رنگ، نرخ مصرف الکل و خودکشی بالایی دارند، یا میزان نرخ خودکشی آنها کمی کمتر از حد میانگین است. میتوان نتیجه گرفت نرخ مصرف الکل با نرخ خودکشی رابطه مستقیم دارد و برای برخی کشورها این رابطه پررنگتر و برای برخی کشورها این رابطه کمرنگتر است.

. در نهایت از پنجره preview نمای نهایی گراف را به دست می آوریم. شکل ۲۴ preview گراف را نشان میدهد.



شکل ۲۴- preview گراف

۷- نتیجه گیری

- در سوال اول که مربوط به اطلاعات یک خرده فروشی وسایل نقلیه می‌باشد به دنبال یافتن این بودیم که ترکیبی از محصولات را در چه کشوری به فروش برسانیم که فروش بیشتری نسبت به سایر موارد داشته باشد. همچنین در سناریوی اول به دنبال کشف مشتریان واقعی هر دسته از محصولات در کشورهای مختلف بودیم. طبق تحلیل شبکه این سناریو دریافتیم که ۳ دسته وسایل نقلیه پرفروشتر هستند که به ترتیب اتوموبیلهای کلاسیک، موتور سیکلتها و اتوموبیل های وینتج می‌باشند. اتوموبیلهای کلاسیک با اختلاف بسیار زیادی در رتبه پرفروشترین وسایل نقلیه قرار گرفته است. کشورهای آمریکا و فنلاند برای فروش اتوموبیلهای کلاسیک بهترین گزینه می‌باشند. کشورهای فرانسه و آمریکا برای فروش موتور سیکلت و کشور اسپانیا برای فروش اتوموبیلهای وینتج بهترین گزینه می‌باشند. با توجه به حجم سفارشها در هر کشور میتوان دریافت سفارشهای اتوموبیل کلاسیک و موتور سیکلت در آمریکا حجم سفارشات بیشتری نسبت به دیگر کشورها داشته است. در آخر میتوان نتیجه گرفت آمریکا بهترین مکان برای سرمایه گذاری و خرید و فروش وسایل نقلیه با حجم بالا می‌باشد. نتایج پرفروشترین اجناس به همراه اطلاعات سفارش و شخص سفارش دهنده در جدول ۶ آورده شده است. لازم به ذکر است اطلاعات این مجموعه داده مربوط به سال ۲۰۰۵ می‌باشد و ممکن است نتایج در سالهای بعدی متفاوت باشند.

جدول ۶- اطلاعات پرفروشترین اجناس

نام وسیله نقلیه	شماره سفارش	کشور	نام سفارش دهنده	شماره تلفن سفارش دهنده
اتوموبیل کلاسیک	۱۰۴۰۰	آمریکا	The Sharp Gifts Warehouse	4085553659
اتوموبیل کلاسیک	۱۰۴۱۳	آمریکا	Gift Depot Inc.	2035552570
اتوموبیل کلاسیک	۱۰۳۸۱	آمریکا	Corporate Gift Ideas Co.	6505551386
موتور سیکلت	۱۰۳۸۸	آمریکا	FunGiftIdeas.com	5085552555
موتور سیکلت	۱۰۳۶۲	آمریکا	Technics Stores Inc.	6505556809
موتور سیکلت	۱۰۴۰۲	فرانسه	Auto Canal Petit	(1) 47.55.6555
اتوموبیل وینتج	۱۰۳۸۵	آمریکا	Mini Gifts Distributors Ltd.	4155551450
اتوموبیل وینتج	۱۰۳۷۹	اسپانیا	Euro Shopping Channel	(91) 555 94 44

- در سوال دوم که مربوط به اطلاعات حوزه سلامت کشورهای جهان می‌باشد به دنبال این بودیم که ترکیبی از نیروهای درمانی شامل پزشک، دندانپزشک، ماما و داروساز را در چه کشورهایی قرار دهیم تا مطمئن شویم افراد آن منطقه سرویسهای درمانی را دریافت میکنند. لازم به ذکر است کشورهایی که در این سوال مورد بررسی قرار گرفته اند تعداد نیروی کادر درمانیشان کمتر از حد معمول بوده است. پس از تحلیل و آنالیز گراف به این نتیجه دست پیدا کردیم که کلیه کشورهای مورد بررسی، در ۳ جامعه یا کلاس قرار میگیرند.

دسته اول شامل کشورهاییست که نیازمند به ۲ تا ۴ گروه از نیروهای درمانی هستند. کشورهایی مانند عراق، مصر، پاکستان، سنگال، اندونزی و آفریقای جنوبی. نیازمندی این کشورها عموما به ماما، پزشک و دندانپزشک می‌باشد.

دسته دوم شامل کشورهاییست که نیازمند به ۲ گروه از نیروهای درمانی هستند، مانند چین، ایران، ترکیه، کره جنوبی و مالزی. نیازمندی این کشورها عموما به ماما و دندانپزشک می‌باشد.

دسته سوم شامل کشورهاییست که نیازمند به ۱ گروه از نیروهای درمانی هستند. کشورهایی مانند کانادا، سوئد، امارات، اسرائیل و انگلستان. نیازمندی این کشورها عموما به ماما می‌باشد.

در انتها میتوان نتیجه گرفت بیشترین نیازمندی به گروه ماما و کمترین نیازمندی به گروه داروساز می‌باشد. همچنین قاره های آسیا و آفریقا نیازمندی بیشتر به خدمات درمانی دارند و کشورهای اروپایی نیازمندی کمتری

به نیروهای درمانی دارند. کشورهایی که به هر ۴ گروه از خدمات درمانی نیاز دارند شامل بنین (کشوری در غرب آفریقا)، بورکینافاسو (کشوری در غرب آفریقا)، گواتمالا (کشوری در آمریکای مرکزی) و مالاوی (جنوب شرق آفریقا) می‌باشند. لازم به ذکر است این مجموعه داده مربوط به سالهای ۲۰۱۷ و ۲۰۱۸ بوده و ممکن است نتایج در حال حاضر متفاوت باشد.

- در سوال سوم که مربوط به اطلاعات حوزه سلامت کشورهای جهان می‌باشد به دنبال این بودیم که مصرف الکل در یک کشور چه تاثیری در خودکشی افراد آن کشور دارد؟ لازم به ذکر است این بررسی برای افراد بالای ۱۵ سال در هر کشور انجام میشود. طبق تحلیل شبکه دریافتیم که کشورها به ۲ جامعه یا گروه تقسیم میشوند. گروه اول شامل کشورهاییست که نرخ مصرف الکل و نرخ خودکشیشان کمتر از میانگین است، در این جامعه میتوان گفت خودکشی و تاثیر الکل رابطه مستقیمی دارند.

جامعه دوم شامل کشورهاییست که نرخ مصرف الکل و خودکشیشان بالاتر از حد میانگین است و در برخی از کشورها نیز نرخ خودکشی اندکی کمتر از حد میانگین است. میتوان گفت در این کشورها نرخ بالای خودکشی مربوط به نرخ مصرف الکل بالاست و رابطه تقریبا مستقیمی بین این دو وجود دارد.

در حالت کلی میبایست به این نکته توجه کرد که خودکشی به عوامل بسیار زیاد دیگر نیز بستگی دارد و مصرف الکل میتواند یکی از عوامل آن باشد.