Cayley’s Formula Applied to Social Networks

Discrete Mathematics (MAT 301) Fall 2019

Douglas DeBettencourt, Gina Pantano, Lucas Repke

May 23, 2021

Background

Arthur Cayley was a British mathematician born at Richmond in Surrey, England in August 1821. Although he was born in England, Cayley spent the first seven years of his life in St. Petersburg, Russia where he was raised by his father Henry Cayley and mother Maria Antonia Doughty [3]. Cayley always had a natural gift for mathematics. He began school at Trinity College, Cambridge at the early age of 17. Cayley finished his undergraduate degree as Senior Wrangler and won the first Smith prize. Senior Wrangler is the top mathematics undergraduate student at Cambridge University. The Smith prize are two prizes awarded to two outstanding research students in mathematics and theoretical physics, which are now known as the Smith-Knight Prize and Rayleigh-Knight Prize. Cayley went on to attend Lincoln's Inn to become a lawyer over the next fourteen years. Throughout his law profession, Cayley still published about 250 mathematical papers [3]. In 1863, Cayley was appointed Sadleirian professor of Pure Mathematics at Cambridge. He published over 900 papers by the end of his career covering nearly every topic of modern mathematics [4].

Cayley helped develop the theory of algebraic invariance, helped develop the theory/algebra of matrices, worked on non-euclidean geometry, studied analytical geometry of n-dimensions, published a treatise on elliptic functions, and was the president of the British Association for the Advancement of Science [4]. To show the impact of his research, O'Conner and Robertson from the University of St. Andrews, Scotland state that, "his development of n-dimensional geometry has been applied in physics to the study of the space-time continuum, and his work on matrices served as a foundation for quantum mechanics, which was developed by Werner Heisenberg in 1925" [4]. During his study of n-dimensional geometry, Cayley made several contributions to graph theory. In particular Cayley solved the problem of how many different trees one can construct with n labeled vertices, which is also known as Cayley's Formula. This mathematical formula made significant contributions to graph theory by its various applications.

One application of this formula is with social networking. "A social network is a graph model for the interactions among the members of a group. Ideas from graph theory apply directly to such models, and sociological concerns raise interesting mathematical questions about graphs" [2]. A social network graphs vertices are called actors and edges are called ties, but we will only refer to them as vertices and edges throughout the paper. Although this paper will primarily focus on Cayley's formulas applications to social networking, his formula can be applied to several kinds of networks. Cayley's formula can be applied to economic networks to show relations between organizations, to informational networks for navigation, to technological networks for routers/transportation, and to biological networks to represent biological systems. Cayley's formula essentially calculates the different number of relations one can have for each of these networks. We will go into more detail regarding the formulas application in our conclusion.

Definitions and Cayley's Formula

Before proving Cayley's Formula, it is important to lay down the background and foundation in order to fully understand the concepts that will be discussed. There are multiple different terms that must be defined in order to fully understand the future explanations of Cayley's Formula. These terms will also be crucial to understanding the proof of Cayley's Formula as well. Here are the essential definitions:

- **Graph**: a visual representation of relations between objects, a graph G = (V, E), a pair of two sets, set V of vertices and set E of edges which connect the vertices [2]
- **Adjacent Vertices**: two vertices that are connected by an edge [2]
- **Connected Graph**: a graph G such that there is a path between any pair of vertices in G [2]
- **Disconnected Graph**: a graph which has multiple disconnected components [2]

- **Simple Graph**: an undirected graph with no loops and no multiple edges [2]
- **Circuit**: a path in a graph G that begins and ends at the same vertex [2]
- **Cycle**: a circuit that does not repeat vertices [2]
- **Degree**: the number of edges that have the vertex 'v' as an endpoint [2]
- **Tree**: a connected undirected graph with no circuits [2]
- **Forest**: a collection of unconnected trees [2]
- **Labeled Tree**: a tree in which each vertice is assigned a number, 1 to 'n', with 'n' being the number of vertices in the tree [2]

These are the definitions that will be essential to understanding future explanations and proofs that will be laid out for Cayley's formula.

Cayley's Formula

Now that the definitions have been laid out, we can begin to discuss Cayley's Formula and the different methods of proofs that go along with it. Cayley's Formula gives us the number of different trees that can be constructed on n vertices [1]. This process can be thought of in two ways. One way is to start with n vertices and to place edges on them until a tree has been made [1]. Another way is to think of it as starting with a complete graph, K_n, and removing edges from it to again make a tree. Cayley's Formula is a formula that tells us how many different ways we can do these processes. This is important as when n gets too large, counting each possibility manually would be nearly impossible, thus Cayley's Formula gives us a solution to find the number of trees on n vertices without having to do this. The number of different ways to make a tree on n vertices are called spanning trees on n vertices. Spanning trees on n vertices will be denoted by T_n. An example of this can be shown using the elements of T_4:

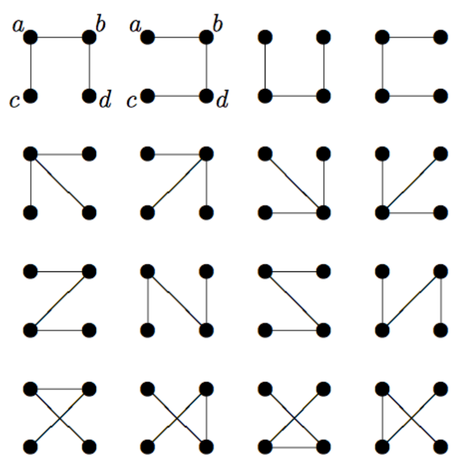

Figure 1: This figure shows all of the possible spanning trees for four labled vertices a, b, c, and d

Cayley's Formula tells us that in T_4, there are 16 different possible trees. The reason there are 16 is because each tree above is actually considered to be different, even though the trees in each row are really just rotations of each other [1]. The reason they count separately is because the vertices are labeled, as is shown in the first two trees in the first row of the figure above. For example, in the two labeled graphs above, a is connected to c in the first graph, but not in the second graph. In the first graph, a, also has a degree of 2, but a only has degree of 1 in the second graph. Since the vertices are labeled, the trees in each row that are rotations actually count as separate trees giving us 16 trees for T_4. These 16 trees could be obtained by either adding edges to 4 vertices, or by taking edges away from a complete graph, K_4 [1]. Cayley's Formula tells us that by doing either of these processes, we will get 16 trees for T_4.

$$T_n = n^{n-2}$$

When applied to T_4, it is easy to see that $|T_4| = 4^{4-2} = 4^2 = 16$ different trees on 4 vertices. We can also apply it to $T_3 = 3^1 = 3$, which makes sense since the only possible trees on 3 vertices would be a 'V' [1]. This shows us that Cayley's Formula works for small n values, but now we will prove that Cayley's Formula works for all n values. Sometimes it is easier to prove something by proving something more general. This is what we will do in the 'A Forest of Trees' proof as we will prove the general formula which can then be applied to the more specific form, Cayley's Formula.

A Forest of Trees

The specific family of forests will be defined as, $T_{n,k}$ [1]. Now, let A be an arbitrary set of k vertices from 1, 2, ..., n. We can now define $T_{n,k}$ as the set of all forests on n vertices with k trees such that each element of A belongs to a different tree [1]. In order to understand the proof and move on, it is important to identify two points; one, a point on 1 vertex by itself is a tree. Second, it is important to note that when counting the elements of $T_{n,k}$, the elements that are in A do not matter, only the amount of elements in A matters which has been defined as k [1]. Now, let us examine $T_{n,k}$:

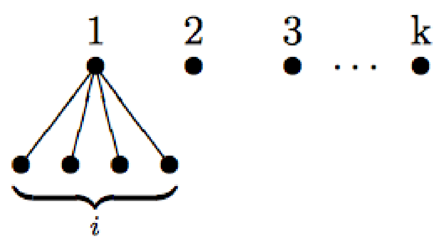

Figure 2:

In Figure 2, we are only examining vertex 1, however it is important to recognize that each vertex (1, 2, ..., k) represents its own individual tree [1]. We can construct $T_{n,k}$ as the following summation:

$$T_{n,k} = \sum_{i=0}^{n-k} \binom{n-k}{i} T_{n-1,(k-1)+i}$$

The summation will range from 0 to $n-k$. The variable i can range from 0 to $n-k$ and Vertex 1 can be adjacent to i of the remaining $n-k$ vertices [1]. Vertex 1 could now be removed from the previous graph so that we would get $(k-1)+(i)$ vertices that have to be apart of separate trees. Now, across all values of i, we can sum the number of trees we would get from deleting Vertex 1 [1]. This is where our summation formula above comes from and how it is created. Lastly, before our proof, it is important to identify the values for which the summation is not defined for:

$$T_{0,0} = 1$$
$$T_{n,0} = 0$$

which implies,

$$T_{n,n} = 1$$

Proposition:

$$T_{n,k} = kn^{n-k-1}$$

Proof. We want to prove our proposition (Equation 2) using Equation (1) and induction:

$$T_{n,k} = \sum_{i=0}^{n-k} \binom{n-k}{i} T_{n-1,(k-1)+i} \tag{1}$$

$$T_{n,k} = \sum_{i=0}^{n-k} \binom{n-k}{i} (k-1+i)(n-1)^{(n-1)-(k-1+i)-1} \tag{2}$$

Switch the order of the summation letting $i = (n-k) - i$,

$$\begin{aligned}
T_{n,k} &= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1-i)(n-1)^{i-1} \\
T_{n,k} &= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1)^{i} - \sum_{i=1}^{n-k} \binom{n-k}{i} i(n-1)^{i-1} \\
T_{n,k} &= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1)^{i}(1)^{n-k-i} - \sum_{i=1}^{n-k} \frac{(n-k)! \cdot i}{(i)!(n-k-i)!} \cdot (n-1)^{i-1} \\
T_{n,k} &= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1)^{i}(1)^{n-k-i} - (n-k) \sum_{i=0}^{n-k} \binom{n-k-1}{i-1} (n-1)^{i-1} \\
T_{n,k} &= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1)^{i}(1)^{n-k-i} \sum_{i=0}^{n-k-1} \binom{n-k-1}{i} (n-1)^{i}(1)^{n-k-1-i} \\
T_{n,k} &= (n-1+1)^{n-k} - (n-k)(n-1+1)^{n-k-1} \\
T_{n,k} &= n^{n-k} - n^{n-k} + kn^{n-k-1} \\
&\boxed{T_{n,k} = kn^{n-k-1}}
\end{aligned}$$

Q.E.D [1]

When $k = 1$, the formula reduces to Cayley's formula which states the total number of single spanning trees one can construct with n labeled vertices.

$$T_{n,k} = kn^{n-k-1}$$
$$T_{n,1} = (1)n^{n-1-1}$$
$$\boxed{T_{n,1} = n^{n-2}}$$

Discussion

	k									
n	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	1	1	0.75	0.5	0.3125	0.1875	0.1094	0.0625	0.0352	0.01953
3	3	2	1	0.4444	0.1852	0.0741	0.0288	0.0110	0.0041	0.00152
4	16	8	3	1	0.3125	0.0938	0.0273	0.0078	0.0022	0.00061
5	125	50	15	4	1	0.2400	0.0560	0.0128	0.0029	0.00064
6	1296	432	108	24	5	1	0.1944	0.0370	0.0069	0.00129
7	16807	4802	1029	196	35	6	1	0.1633	0.0262	0.00416
8	262144	65536	12288	2048	320	48	7	1	0.1406	0.01953
9	4782969	1062882	177147	26244	3645	486	63	8	1	0.12346
10	100000000	20000000	3000000	400000	50000	6000	700	80	9	1

Figure 3: We created this graph as a visual representation of the number of single spanning trees one can construct for increasing n labeled vertices.

Some trends within Cayley's Formula on a scale larger than $k = 1$ can reinforce the validity of the formula within a larger context of problems. Furthermore, if $n = k$ then only a single tree could be found within those vertices. Shown by algebra below.

$$T_{k,k} = \cancel{k}\cancel{(k)^{-1}} = 1$$

For all $n - 1 = k$, the value of $T_{n,k} = n - 1$. Which also makes sense because of a tree of k vertices with one more vertex in the set of vertices means n different trees could be created because any given vertex, from 1 to n could be not included in the set of used vertices.

$$T_{n,n-1} = (n-1)n^{\cancel{n}-\cancel{n}+\cancel{1}-\cancel{1}} = n - 1$$

Notice that for n=1, $T_{n,k} = k$. This makes sense because over 1 vertex, any number k trees could be found given k labeled vertices. For example, given the points a,b,c choose only one of them to create a tree of 1 vertex then 3 different trees could be considered at one time depending on which point, a,b,c is viewed. For every k such that $1 < k < n$, the value is not an integer indicating a complete tree cannot be found.

$$T_{n,k} = kn^{n-k-1}$$
$$T_{1,k} = k\cancel{(1)^{-k}} = k$$

Notice the section of the graph with non-integer values in Figure 3. This makes sense again because it is logically impossible to find a tree with $n + a$ $(a \epsilon Z, a > 0)$ vertices within a set of n vertices. Every value will be some number between 0 and 1 which cannot be an integer, because it will be numerically between two other consecutive integers. Since a non-integer number of trees is a contradiction by basic counting principles, every case within these parameters for n and k would make sense as values supporting the validity of Cayley's formula in a larger context.

$$T_{n,n+a} = (n+a)n^{-(a+1)}$$

Since $n + a < n^{a+1}$ for all a,

$$0 < \frac{n+a}{n^{a+1}} < 1$$

Conclusion

Cayley's formula can be easily applied to social networking because it is an analysis of graph theory. Within graph theory, edges represent relationships (ties) between people and vertices represent people (actors). Cayley's formula specifically refers to trees. Trees are very common ways to represent familial relationships and phylogenies. Phylogenies are trees representing common ancestry among organisms, often used to display the long term effects of evolution. The application of Cayley's formula within familial trees is significant because it can show how many different ways a group can be related to each other. Cayley's formula shows this in relation to labeled trees, which works because the family trees are labeled by each person within the family. The generic case for Cayleys formula can be used to describe how many ways a certain group of people can be related to each other within a familial relationship. A similar idea could be used to apply this to a group of people considering mentors and subordinates. An easy way to track how far apart two people are in some field of study could indicate why they have similar views on a particular topic. For example, if two people have 5 edges between them while another set has two edges between them, then the pair with 5 edges can be a more reliable source because the topical views have been reinforced through more times teaching and learning the material. Cayley's formula could be further refined to determine the average number of connections between any two points on the tree of n vertices, making an analysis of an entire scholarly field possible.

References

[1] Casarotto, Chad. 2006. *Graph Theory and Cayley's Formula*.
Retrived from `http://www.math.uchicago.edu/ may/VIGRE/VIGRE2006/PAPERS/Casarotto.pdf`

[2] Hunter, David J. 2012. *Essentials of Discrete Mathematics*, 2nd Edition. Sudbury, MA. Jones and Bartlet Learning

[3] O'Connor, J. J. and Robertson, E. F. 2014. *Arthur Cayley*. School of Mathematics and Statistics, University of St. Andrews, Scotland.
`http://mathshistory.st-andrews.ac.uk/Biographies/Cayley.html`

[4] MacFarlane, Alexander. 1916. *Chapter 5: 'Arthur Cayley', Lectures on Ten British Mathematicians*. London, England: Chapman and Hall. Lit2GO.
`https://etc.usf.edu/lit2go/27/lectures-on-ten-british-mathematicians/274/chapter-5-arthur-cayley/`