Ben Levine & Yegor Tkachenko

Leskovec, Adamic, Huberman (2008) – The Dynamics of Viral Marketing

Background

- To date, no academic research identifying "systematic patterns about the nature of knowledge-sharing and persuasion by influencers and reactions by recipients of online social networks"
- Fat tail of sales of products online – lots of sales come from obscure products
    - Top 20% of products only contribute half the sales (compare to 80/20 rule)
    - This has implications for marketing – tough to apply traditional marketing strategies given wide array of niche products
- *Viral marketing:* "Diffusion of information about a product and its adoption over a network"
- eWOM research still nascent, findings are stylized facts rather than coherent theory
    - "Most of the previous research on the flow of information and influence through networks has been done in the context of epidemiology and the spread of diseases throughout a network"
    - *SIR/SIRS models:* susceptible - infected - recovered - (susceptible again)
        - *Epidemic threshold:* given a set of initially infected nodes (people who bought the product without recommendations), conditions under which disease will either dominate or die out
    - *Bass diffusion models:* rate of adoption is a function of the current proportion of the population that have already adopted
    - *Diffusion equation models:* S-shaped curve of cumulative adoption – can accurately describe WOM product diffusion in aggregate, but not at the level of the individual person, which is one of this paper's contributions
    - *Threshold models:* A node adopts if the sum of the connection weights is greater than a certain threshold (b/t 0 and 1, sourced from a probability distribution). Connection weights vary across neighbors and nodes.
    - *Cascade models:* Every time a neighbor adopts, there is some probability that the node will adopt. Probabilities vary across neighbors and nodes
    - But, these models are based on assumed vs. *measured* influence effects
    - Previous research modeled adoption of one product; this is the first paper to *measure* and model *many products simultaneously*

Data Description

Recommendation referral program by large retailer; after every purchase, option is given to buyer to recommend their purchase to friends via email. The first recommendee to purchase the

recommended product receives a 10% discount along with the recommender, but any other recommendee purchaser receives no discount.

- 15.6M recommendations from 4M unique users, 500K products (4 product groups)
    - Collected 2001-2003
    - Limitations:
        - Recommendees could purchase elsewhere
        - Users may consider these emails spam or doubt recommender's motives
        - Emails sent before recommender receives product; therefore, cannot provide direct testimonial
- Directed multi graph. The nodes represent customers, and a directed edge contains all the information about the recommendation.
- The edge (*i, j, p, t*) indicates that *i* recommended product *p* to customer *j* at time *t*.
    - Note that as there can be multiple recommendations of between the persons (even on the same product) there can be multiple edges between two nodes.
- *Buy-bit*
    - Action recorded when first person buys a recommended item
- *Buy-edge*
    - Action recorded when someone recommends an item they have been recommended
    - Limitation: it is possible for consumer to not be first to buy and also not recommend product to others, in which case this is not recorded

Group	n_c	r_c	e_c	b_{bc}	b_{ec}
Book	53,681	933,988	184,188	1,919	1,921
DVD	39,699	6,903,087	442,747	6,199	41,744
Music	22,044	295,543	82,844	348	456
Video	4,964	23,555	15,331	2	74
Full network	100,460	8,283,753	521,803	8,468	44,195

Table 2: Statistics for the largest connected component of each product group. n_c: number of nodes in largest connected component, r_c: number recommendations in the component, e_c: number of edges in the component, b_{bc}: number of buy bits, b_{ec}: number of buy edges in the largest connected component, and b_{bc} and b_{ec} are the number of purchase through a buy-bit and a buy-edge, respectively.

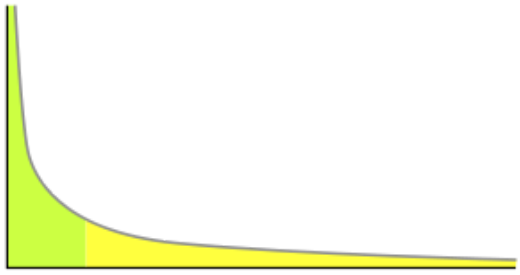
- *Communities/components*
    - Nodes connected by other nodes (directly and indirectly)
- How do communities grow?
    - 75/25 recommendation comes from node in the largest component/smaller component joins larger one

- "Six degrees of separation" principle does not hold
    - Largest connected component comprises only 2.5% of all nodes

Key concepts

Power law

Power law is a relationship, where one quantity varies as a power of another. For instance, considering the area of a square in terms of the length of its side, if the length is doubled, the area is multiplied by a factor of four.



If vertical axis is sales, and horizontal axis are items ranked by sales, this plot could demonstrate the 80%-20% rule, where 80% of items generate 20% of sales (long tail), and vice versa. (Note, in this paper top 20% of items contribute ~50% of sales).

Functional form: $f(x) = \alpha x^{-k}$
Equivalently, in form estimable with OLS: $\log(f(x)) = \log(\alpha) - k\log(x)$
k is often called a power-law exponent and its estimate is used to characterize the distribution.

Scale invariance - scaling input by a factor scales function itself proportionately.
$f(cx) = \alpha(cx)^{-k} = c^{-k}f(x) \propto f(x)$

Cascades

Cascade in general means that some event triggers future events, those events trigger more future events, and so on - across connected nodes in a graph. The exact definition varies across applications.

In this paper, cascade is a network consisting of customers (nodes) who purchased the same product as a result of each other's recommendations (edges). The graph is constructed to be *temporally causal* - for each node (customer), all incoming edges (recommendations) for some

product occur before all outgoing edges for that product. Moreover, all incoming edges (recommendations) that occur after the first purchase by the node are removed (ignored).

Cascade sizes generally follow a power law distribution. For example, for DVDs, the distribution of cascade sizes follows a power law with k=1.56. It says that a large number of cascades are rather short, but some are quite long.

Propagation model of recommendations

Each recipient forwards a recommendation received at time t with probability p_t (if recommendation value probabilistically exceeds a threshold). Note that p_t is itself random. If N_t is the number of recommendations at step t, then

$$N_{t+1} = p_t N_t .$$

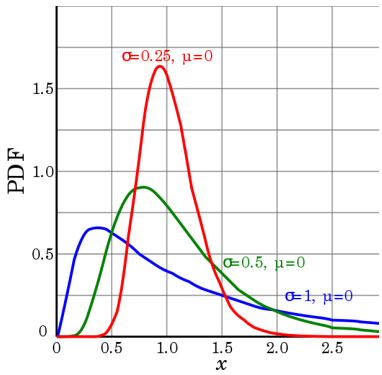
Then we get after subtracting and dividing by N_t,

$$(N_{t+1} - N_t)/N_t = p_t - 1$$

Summing across t for large T, up to a unit constant we have that log of the number of messages

$dN/N = \log(N) = \sum p_t$, which as a sum of random variables, by CLT, follows Normal dist.

In other words, number of messages N is log-normally distributed.



P(N) = logNormal

For large variances (blue line, most skewed to the left), this density describes the situation where small N cascades are more frequent, but large N long cascades are also sometimes observable. Taking log of the normal density expression, for very large variance, we have

$$\log(P(N)) \approx -\log(N) - \log(\sqrt{2\pi\sigma^2})$$

where, the last term is a constant, and so we have exactly the power law expression with exponent k= -1. We get close to this model in the case of cascade sizes of DVDs, but not so close with other product groups (Figure 6).

Connected components

- *Path* in a graph is a finite or infinite sequence of edges which connect a sequence of nodes which, by most definitions, are all distinct from one another.
- *Directed path* is again a sequence of edges which connect a sequence of nodes, but with the added restriction that the edges all be directed in the same direction.
- *Connected component* of an *undirected* graph is a subgraph in which any two nodes (vertices) are connected to each other by paths.
- *Strongly connected component* is a subgraph of a *directed* graph where each pair of nodes is contained within oppositely oriented directed paths.
- *Weakly connected component* is a maximal subgraph of a directed graph such that for every pair of vertices **u, v** in the subgraph, there is an undirected path from **u** to **v** if directions were removed.

The analysis is very often performed using the largest (weakly) connected component of a graph. This component usually contains a large portion of events of interest.

Community detection

Community finding algorithm the authors use breaks up a weakly connected component into parts such that Q is maximized, where

$$Q = (\#\ of\ edges\ within\ community) - (expected\ number\ of\ edges)\ .$$

Communities of specialized interests can be identified as follows. If p_c is a proportion of all recommendations that belong to category of interests c, and x_g is the number of recommendations sent by the community, we have the expected number of recommendations in a given category to be purely by chance within $p_c * x_g \pm \sqrt{p_c * (1 - p_c) * x_g}$. When the actual number exceeds this interval, we can mark the community as having that special interest.

Stylized Facts

- Additional purchases that resulted from recommendations due to campaign were "a drop in the bucket of sales that occur through the website".
- Probability of purchasing a product increases with the number of recommendations received, but quickly saturates to some low probability.
- People who send out many recommendations tend to be less influential in getting recipients to pursue recommendation.
- Nice interest-based community structure can be recovered around most frequently recommended items.
- Smaller communities tend to be more conducive to viral marketing.
- Higher product price increases the probability that a recommendation will be accepted.

	$\log(s)$	$\log(n)$	$\log(n_s)$	$\log(n_e)$	$\log(r)$	$\log(e)$	$\log(p)$	$\log(v)$	$\log(t)$
$\log(s)$	1								
$\log(n)$	0.275	1							
$\log(n_s)$	0.103	0.907	1						
$\log(n_r)$	0.310	0.994	0.864	1.000					
$\log(r)$	0.396	0.979	0.828	0.988	1				
$\log(e)$	0.392	0.981	0.831	0.990	0.999	1			
$\log(p)$	0.185	0.098	0.088	0.098	0.107	0.106	1		
$\log(v)$	-0.050	0.465	0.490	0.449	0.421	0.423	-0.053	1	
$\log(t)$	-0.031	0.064	0.071	0.061	0.056	0.056	-0.019	0.269	1

Table 7: Pairwise Correlation Matrix of the Books and DVD Product Attributes. $\log(s)$: log recommendation success rate, $\log(n)$: log number of nodes, $\log(n_s)$: log number of senders of recommendations, $\log(n_r)$: log number of receivers, $\log(r)$: log number of recommendations, $\log(e)$: log number of edges, $\log(p)$: log price, $\log(v)$: log number of reviews, $\log(t)$: log average rating.

Variable	Books Coefficient β_i	DVD Coefficient β_i
const	1.317 (0.0038) **	0.929 (0.0100) **
n	-0.579 (0.0060) **	0.171 (0.0124) **
n_s	0.144 (0.0018) **	-0.070 (0.0023) **
n_r	-0.006 (0.0064)	-0.360 (0.0104) **
r	0.062 (0.0084) **	-0.002 (0.0083)
e	0.383 (0.0106) **	0.251 (0.0088) **
p	0.013 (0.0003) **	0.007 (0.0016) **
v	-0.003 (0.0001) **	-0.003 (0.0006) **
t	-0.001 (0.0006) *	0.000 (0.0009)
R^2	0.30	0.81

Table 8: Regression Using the Log of the Recommendation Success Rate log(s), as the Dependent Variable for Books and DVDs separately. For each coefficient we provide the standard error and the statistical significance level (**:0.001, *:0.1). We fit separate models for books and DVDs.