**Growth accounting: data science techniques to track your user base and profitability**

*Introduction*

The core focus of most companies is growing and monetizing their user base. Doing this intelligently is the aim of a corner of data science often referred to as growth analytics. With the associated concepts (e.g., viral growth, customer retention, lifetime value), companies are able to perform the vital work of tracking and modeling changes to their total number of users and the associated cash flows.

Internet and businesses-to-consumer (“B2C”) companies, whose potential customer base includes the entire internet-connected world ([4.5B+ billion people](t.ly/1VZm) as of 2020) are particularly well-suited to benefit from such analysis. Their customers can easily sign onto a service, and their engagement with particular product features serves as an important measure of business health. These organizations include platforms hosting user-generated content alongside embedded advertisements (Facebook, Twitter, TikTok), logistics companies physically delivering items ordered online (Amazon, Uber, Instacart), and financial technology firms involved in transactions processing (PayPal, Square, Stripe).

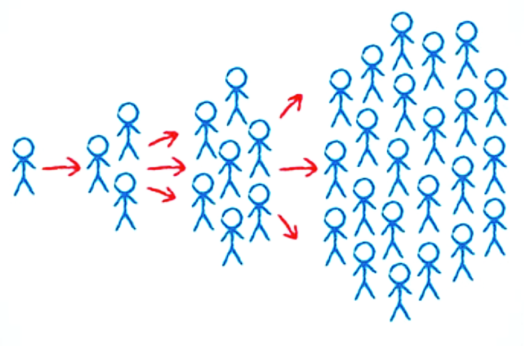
On the other hand, growth analytics tools may be of limited use to companies whose core products are not online do not have the benefit of easily monitoring user engagement. Similarly, business-to-business (“B2B”) companies typically court a small number of prospective customers and often customize their product and pricing to each one.

This article aims to summarize key concepts in growth data science taken from both text ([Lean Analytics](t.ly/JgJYm), [Amplitude](t.ly/P9qB), [Mixpanel](https://mixpanel.com/topics/), [Social Capital](https://medium.com/swlh/diligence-at-social-capital-part-1-accounting-for-user-growth-4a8a449fddfc#.bvu0we2z3), [Tribe Capital](https://tribecap.co/a-quantitative-approach-to-product-market-fit/)) and video ([Alex Schultz](t.ly/Q6G3), [Elliot Shmukler](t.ly/3pZZ5), [Michael Seibel](https://www.youtube.com/watch?v=C27RVio2rOs)) resources.

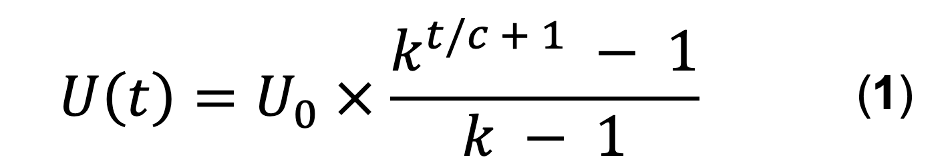
***I. Growth***

*1. Viral growth model*

Borrowed from epidemiology, the [viral growth model](https://readwrite.com/2015/10/09/virality-math-formula/#:~:text=To%20fix%20this%2C%20we%20simply,and%20the%20cycle%20time%20%E2%80%93%20Ct.) allows companies to project the size of their user base over time under the assumption that new users are brought in by existing users.



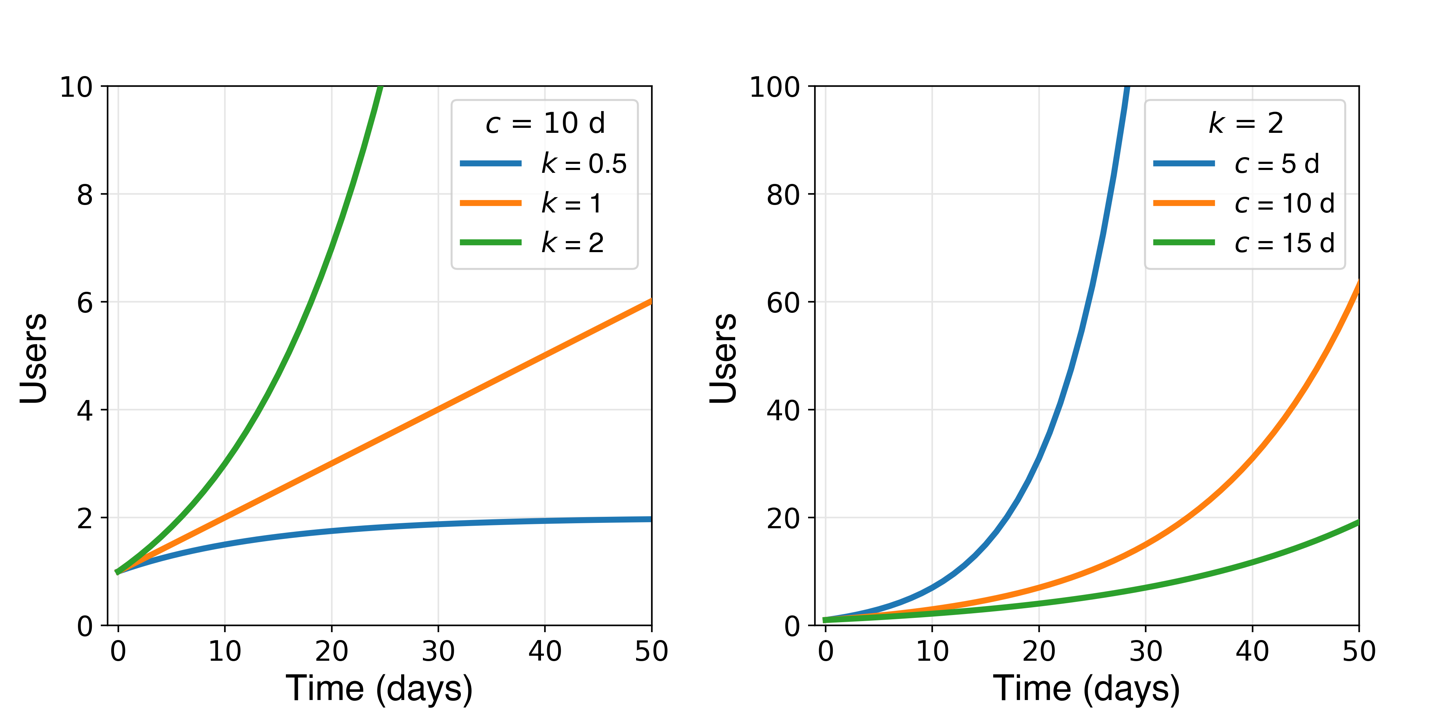
Specifically, the number of users at any particular time *U*(*t*) can be expressed as



where *U*0 is the starting user count, *t*/*c* is the number of cycles elapsed (for time elapsed *t* and cycle time *c*), and *k* is the viral coefficient, or number of new users brought on per existing user.

A. Viral coefficient, *k*

The key assumption of the viral growth model is that new users engage with the product as a result of interaction with current users. Practically speaking, a company can empirically determine its viral coefficient *k* by multiplying the invitation rate of a current user by the acceptance rate of an invited user. For example, if existing users send on average 5 invites, and 40% of invitees accept, the viral coefficient is 5 × 0.4 = 2. In this scenario, each current user will bring on 2 new users. A key goal for many young companies attaining a viral coefficient of one, since *k* = 1 defines the boundary between saturating and accelerating growth (Figure 1, left).



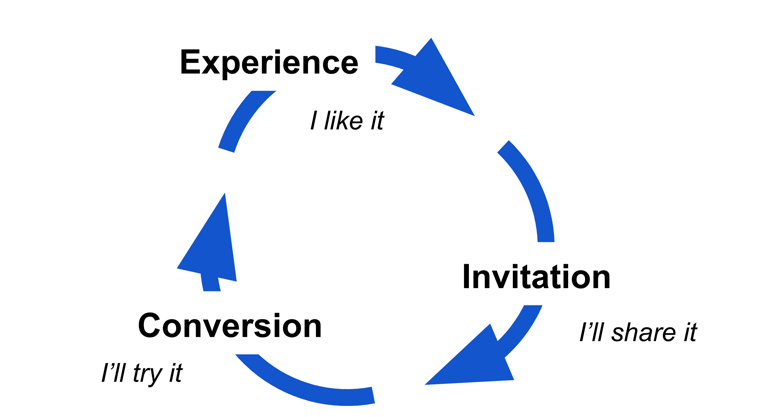
**Figure 1**. Effect of viral coefficient *k* and cycle time *c* on growth of user base according to Equation 1. (Left): linear growth results for *k* ≈ 1 (note: *U* is not defined at *k* = 1) while *k* > 1 and *k* < 1 yield exponential growth and saturation, respectively. (Right) shortening cycle time increases growth rate.

B. Cycle time, *c*

To bring time into the analysis, viral coefficient *k* is raised to power *t*/*c* + 1. Here, cycle time *c* is the time interval separating conversion of one set of users and conversion of their invitees, and *t/c* is thus the number of viral cycles elapsed. Starting with 1 user and a viral coefficient of 2, for instance, the number of users after 1, 2, and 3 cycles will be 3, 7, and 15, respectively. In the first cycle, the first user brought on 2 new users. In the second cycle, those 2 latest users each brought on 2 (4 total) new users. In the third cycle, those 4 latest users each brought on 2 (8 total) new users.

Following this logic, for a given time *t* elapsed, reducing the cycle time *c* is equivalent to increasing the number of elapsed cycles (Figure 1, right). While the viral coefficient dictates the shape of the user growth curve, cycle time controls the compression of the growth curve along the x-direction.

The viral cycle is composed of three steps: (1) a positive user experience is delivered, (2) the user invites others, (3) some of those invitations are accepted (Scheme 1). The cycle time c is can then be calculated as the average time interval separating the conversion of one user and conversion of the user(s) they bring in.

­­­­­­­

**Scheme 1**. Viral cycle.

The viral growth model suffers from a couple of important limitations:

1. Viral coefficient and cycle time will change over time. Ideally, improvements to the product will increase *k* and reduce *c* (see next section), but it’s also possible that market saturation, competition, or unfavorable product changes do the opposite*.*
2. Some new users will discover the product without any invitation. Modeling user growth based on invitations from current users alone will not capture this organic (i.e., self-conversion) component of growth.

As such, viral coefficient and cycle time can be understood as important levers driving growth of the user base, but not necessarily perfect predictors of its size at any point in the future.

*2. Increasing growth rate*

Even if Eq. 1 is an oversimplification, increasing the viral coefficient *k* and reducing cycle time *c* are two critical approaches to increasing user growth.

**Increasing viral coefficient, *k***

Since *k* is the product of invitation rate and acceptance rate, increasing either will increase viral coefficient. To increase ***invitation rate*** (number of invites sent out per existing user, see Figure 2 top), consider doing the following:

1. Present users with an option to invite a list of their contacts pulled from email, social media, etc. This makes it easier to for users to invite others.
2. Build in incentives for sharing. These can be one- or two-way, benefitting the inviter or both inviter and invitee, and can be monetary or otherwise. This makes sending invitations more appealing.
3. Work on extending the user lifetime to allow for more sharing. Once a user drops off, they won’t be inviting potential new users.

On the other hand, to increase ***acceptance rate*** (fraction of invitees who convert, see Figure 2 bottom), you might:

1. Add compelling content to landing page, such as a brief summary of the value proposition to new users.
2. Reduce barriers to new user signup. Requiring as few input fields as possible (e.g., name, email, password) will make it easier for interested prospective users to convert.



**Figure 2.** Increasing viral coefficient by optimizing invitation rate from current users (top: Dropbox example) and acceptance rate from invited prospective new users (bottom: Facebook example).

B. Reducing cycle time, *c*

Reducing cycle time is equivalent to fast-forwarding your growth trajectory, and might be accomplished by simply speeding up each stage of the viral cycle shown in Scheme 1. Consider doing the following:

1. Quickly delivering the key positive user experience (aka. critical event or [magic moment](https://medium.com/egyptian-startup-manual/how-to-conduct-aha-moment-aka-magic-moment-analysis-without-knowledge-of-statistics-or-data-10e59c38ee5)) that brings a new user from evaluation to loyalty mode. Optimize the sequence of navigation choices the user is presented with ([conversion funnel](https://www.hotjar.com/blog/funnel-analysis/)) to get them to this positive experience as fast as possible.
2. Incorporate a “share” feature alongside or shortly after the magic moment so that the satisfied user can invite friends while the positive experience is fresh in their mind. Consider introducing expiring incentives for invitations.
3. Reduce barriers to new user signup. In addition to increasing acceptance rate, requiring as few input fields as possible (e.g., name, email, password) will reduce the time spent on this step.

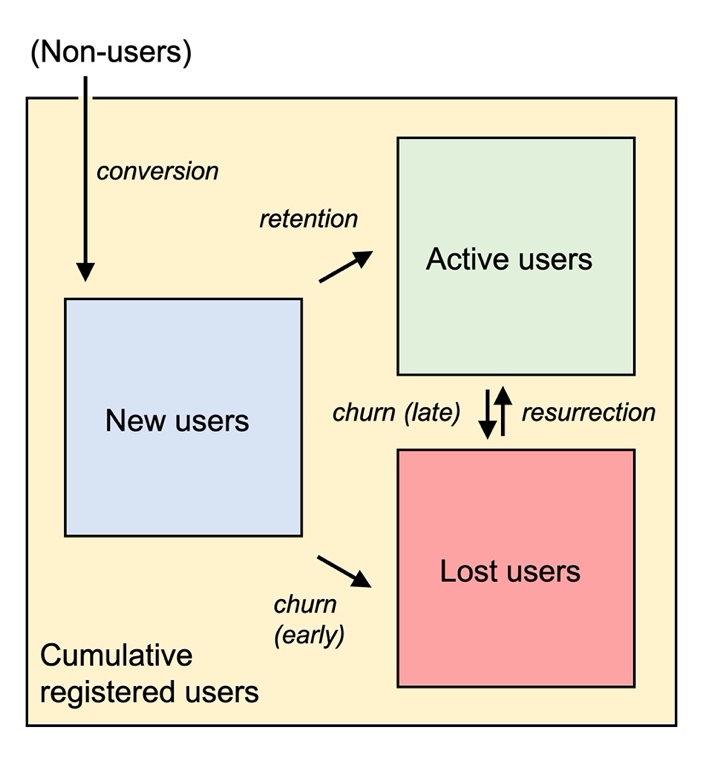
***II. Retention and churn***

While adding new users is essential, it is ultimately *active users* who contribute to revenue by clicking ads or making a purchase. For this reason, it’s important to focus on these users, separating them from the rest of the user base. Indeed, active users are often a small fraction of cumulative registered users: the average mobile app, for instance, [loses 80% of its daily active users](https://andrewchen.co/new-data-shows-why-losing-80-of-your-mobile-users-is-normal-and-that-the-best-apps-do-much-better/) within just 3 days.

*1. Categorizing users*

Most companies can claim as registered users only a tiny fraction of the global internet-connected population. Within this group, there are three smaller subsets (Scheme 2) of users:

1. New users: those who have signed up within the specified time interval
2. Active users: those who have taken a particular action (see next section) within that time
3. Lost users: those who have not taken the action within that time

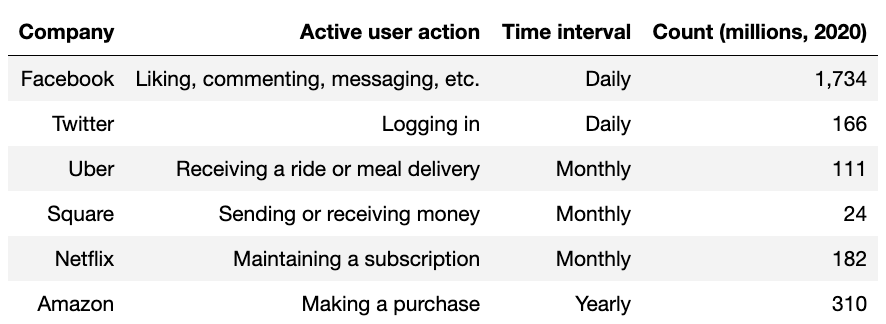


**Scheme 2**. Users within the cumulative registered user base include new, active, and lost (churned) users. New users are either retained as active users or churn due to inactivity. Active users can churn, and churned users can reactivate.

*2. Defining an active user*

An active user is one who has completed some action, ideally one tied to revenue generation, within a particular time interval. Publicly-traded tech companies typically share their active user count and definition of the action that distinguishes active from inactive users in quarterly investor presentations (Table 1).

It’s also necessary to stipulate the time interval in which an action must be taken for the user to be considered active. This is often tied to the frequency with which users are expected to engage with the product. Social media offers up free, personalized content whenever the user has a moment to browse, and many users visit daily. On the other hand, ride-hailing, e-commerce, and mobile payments platforms address a specific need, typically cost money, and are thus used less frequently. Very few people are daily Uber riders, for example, and focusing on that population would amount to ignoring the core customer base.

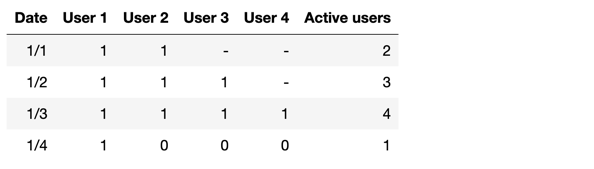


**Table 1**. Active user definition and count for selected companies. See associated references: [Facebook](https://s21.q4cdn.com/399680738/files/doc_financials/2020/q1/Q1-2020-FB-Earnings-Presentation.pdf), [Twitter](https://s22.q4cdn.com/826641620/files/doc_financials/2020/q1/Q1-2020-Shareholder-Letter.pdf), [Amazon](https://etaileast.wbresearch.com/blog/amazons-engaged-buyers-drive-social-media-revenue), [Uber](https://s23.q4cdn.com/407969754/files/doc_financials/2019/sr/InvestorPresentation_2020_Feb13.pdf), [Square](https://s21.q4cdn.com/114365585/files/doc_financials/2019/q4/2019-Q4-Shareholder-Letter-Square.pdf).

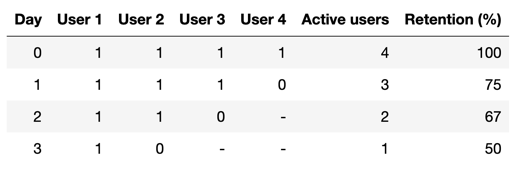
*3. Calculating user retention*

Having chosen an active user definition and time interval relevant to the customer base, it is possible to create an activity log labeling user status over time (Table 2). This can be transformed into a signup-centered activity log by subtracting the date of first use for each user from the date column (Table 3).

Table 3 shows an additional column, the daily retention rate, calculated as the fraction of users who return on a particular day. Temporal resolution of user retention can be adjusted by grouping by week, month, or a custom set of time intervals. Pinterest, for example, looks especially carefully at user retention within [days 1-7 and 28-35](https://jwegan.com/growth-hacking/27-metrics-pinterests-internal-growth-dashboard/).



**Table 2**. Date-centered user activity log displaying user status (active = 1).



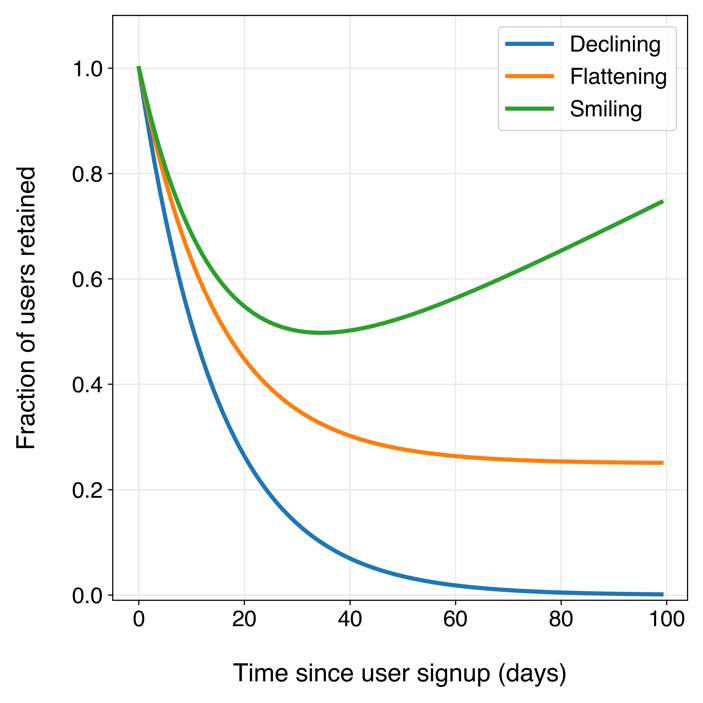
**Table 3**. Signup-centered activity log. The ‘Day’ column refers to days since the user joined. All users are active on day 0 by definition.

The signup-centered formatting (Table 3) has the advantage of surfacing insights specific to milestones along the user journey. For instance, an email sent 3 days after user signup might be expected to boost the number of active users on each user’s day 3. In the date-centered user activity log (Table 2), the associated bump in users would be distributed across dates and harder to pick out. On the other hand, news or product releases that occur on a particular day (for instance, the [#DeleteUber](https://www.businessinsider.com/uber-deleteuber-protest-hundreds-of-thousands-quit-app-2019-4) campaign of 2017) are most visible from a date-centered perspective.

*4. Visualizing user retention*

The signup-centered user activity log can be used to plot a retention curve, which displays the fraction of users retained as active users vs. days since signup. Beyond simply quoting a single value on a particular day, the retention curve offers some temporal resolution, showing how quickly user engagement drops, where it levels off, and if users are resurrected at any point.

Along these lines, there are [three archetypes](https://www.sequoiacap.com/article/retention) for patterns in user retention: declining, flattening, and smiling retention curves (Figure 3).



**Figure 3**. Three classic examples of retention curves.

*Declining retention*

This is the case when the active user base flattens out to zero or near zero in the long term (Figure 3, blue trace), and is typically interpreted as evidence for a lack of product-market fit. Increasing the y-asymptote of the retention curve to some value above zero (i.e., attaining a core group of loyal users) is the most pressing task facing a young company and is related to [solving a critical problem](https://www.youtube.com/watch?v=C27RVio2rOs) that a business or individual is facing.

One notable caveat to this is the video games industry: opportunities for long-term customer retention are [essentially nonexistent](https://www.gamesindustry.biz/articles/2014-10-20-life-is-short-for-mobile-games). One-time smash hits like Farmville and Angry Birds now hold onto a miniscule fraction of their total user base. In such cases, simply slowing the rate of decline to maximize the area under the retention curve (active user days) might be the most reasonable goal.

*Flattening retention*

This is the case when some fraction of new users find value in the product and remain as retained users long after signup (Figure 3, orange trace). Most successful businesses fall into this category. Not all flattening retention curves are the same, however: the y-value where retention flattens out offers a measure of the strength of product-market fit.

Netflix, for example, leads the streaming video segment by this measure, [retaining two-thirds](https://secondmeasure.com/datapoints/netflix-disney-plus-apple-customer-retention/) of their users one year after signup. Because acquiring a new customer is [more expensive](https://hbr.org/2014/10/the-value-of-keeping-the-right-customers) than retaining an existing one, it’s more profitable to grow an active user base via improving retention than by increasing new user signups.

[omit?] Benchmarking: Retention rates also [vary by industry](https://discover.mixpanel.com/rs/461-OYV-624/images/2019-Mixpanel-Product-Benchmarks-Report.pdf): while finance and media products experience, on average, a 7-day retention rate on the order of 10%, e-commerce sees just 1% retention over the same interval.

*Smiling retention*

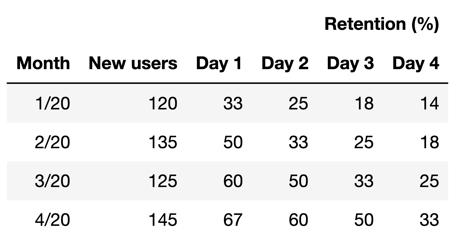
In the ideal case, some fraction of inactive users will return (Figure 3, green trace). Product improvements are typically credited with user resurrection. Adding product categories is one way to win back churned users. Four years after it started selling books online, Amazon started selling [nearly everything else](https://www.nytimes.com/1998/08/05/business/amazoncom-is-expanding-beyond-books.html), prompting inactive book buyers to re-engage through a different set of products.

Network effects also encourage resurrection: as the more drivers signed onto the platform, Uber reduced wait times and thus became more valuable to riders. Similarly, as more smartphone users downloaded WhatsApp, early adopters had more reason to re-engage with the messaging service.

Retention gyrations aside, one thing remains certain: over the longest time horizon, every company (even the most successful ones) eventually loses all of its users. Ten years after claiming 37% share of the US smartphone market, Blackberry phones will [no longer be produced](https://www.businessinsider.com/tcl-killing-blackberry-phones-august-2020-2).

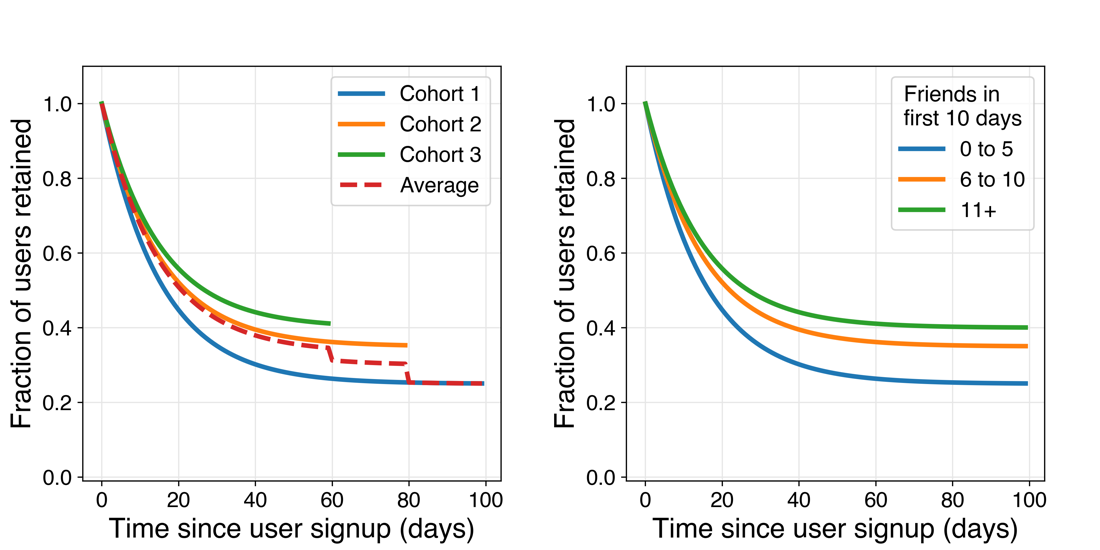
*Segmenting retention data*

It’s often helpful to split up user retention data along some dimension of interest. Grouping users by join date (Table 4) is one common implementation of this concept, and is referred to as [cohort analysis](https://chartio.com/learn/marketing-analytics/cohort-analysis-primer/).



**Table 4**. Retention data aggregated across month of join date.

Visualizing retention data segmented by cohort (Figure 4, left) has the distinct advantage of revealing changes in user engagement across product iterations. In this example, users joining later show a slower decline in engagement and higher long-term retention rate than those who join earlier. While not necessarily evidence of causality, an observed increase user engagement over time can offer some support for the notion that the product is improving, and is not evident by simply looking at averaged retention data.



**Figure 4**.Example retention curves for user groups segmented by join date (left) and by product action (right). Cohort-based retention analysis can confirm improved user engagement with recent product iterations, while action-based retention analysis might inspire future product changes to improve retention. These analyses are complementary and self-reinforcing.

To inspire such product changes, it can be helpful to brainstorm product actions possibly predictive of higher user engagement, and then plot retention curves for users segmented across this axis (Figure 4, right). In a well-known [anecdote](https://mode.com/blog/facebook-aha-moment-simpler-than-you-think/) from Facebook’s early growth team, retention was found to correlate strongly with the number of friends a new user acquires in their first 10 days on the platform. Based on this observation, the team focused on funneling new users towards the critical “add friend” action.

One might also examine retention across user demographic information (e.g., age, gender, geography) or product engagement. Noting differences in retention between demographic groups can offer a call-to-action, perhaps making product changes to address the deficit, or doubling down on marketing to users who fit the loyal profile.

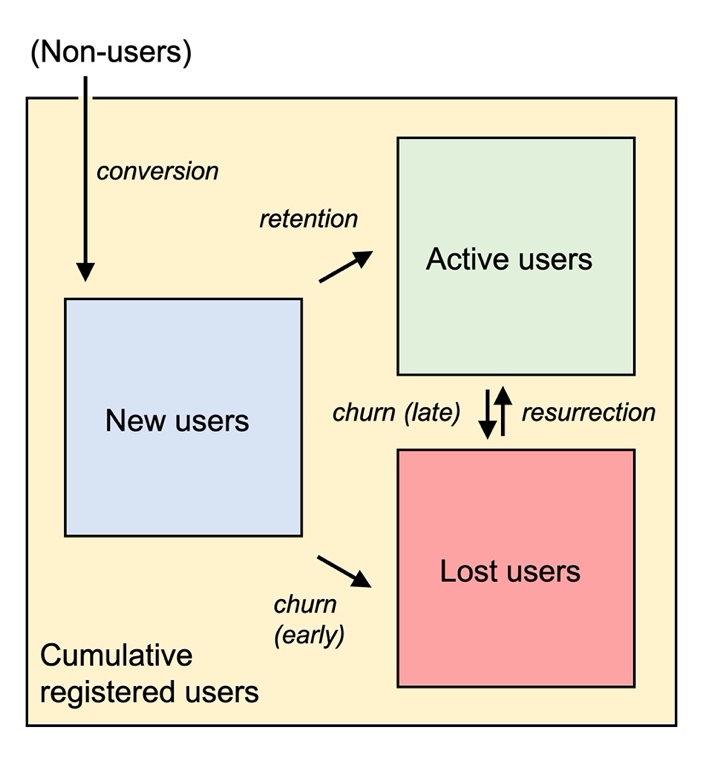
***III. Modeling user growth***

The total user base can be divided into new, active, and lost users (Scheme 2).

As the arrows illustrate, there are 5 possible paths that can be taken:

\*non-users become new users through conversion (either by invites or organic discovery)  
\*new users become active users through retention

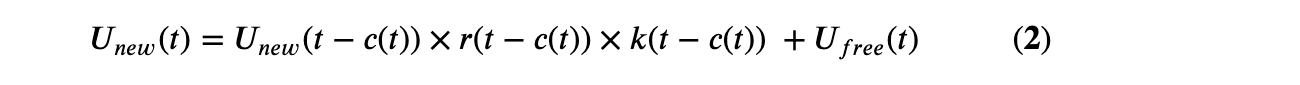
\*new users become lost users through early churn  
\*lost users become active users through resurrection   
\*active users become lost users through late churn



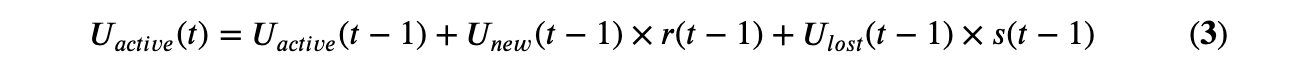
**Scheme 2**. Users within the cumulative registered user base include new, active, and lost (churned) users. New users are either retained as active users or churn due to inactivity. Active users can churn, and churned users can reactivate.

This framework lends itself to calculating the distribution of users at any point in time, for the general case of time-dependent user growth and churn/retention/resurrection rates. Specifically, the number of users falling into each category at time *t*, given as *U*new(*t*), *U*active(*t*), and *U*lost(*t*), can be calculated using the inputs:

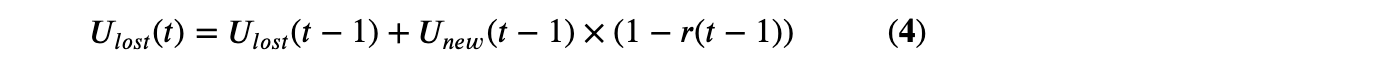
* *k*(*t*) viral coefficient
* *c*(*t*) cycle time
* *r*(*t*) retention rate
* *s*(*t*) resurrection rate
* *U*free(*t*) uninvited new users

**New users** at time 𝑡 is equal to product of new users *U*new, retention rate *r*, and viral coefficient *k* from one cycle ago, plus users who discover the product independently *U*free:  
  


This expression makes the conservative but reasonable assumption that invites only come from retained users. Cycle time at time *t* is *c*(*t*), and the term *t* – *c*(*t*) represents the time at one cycle ago.

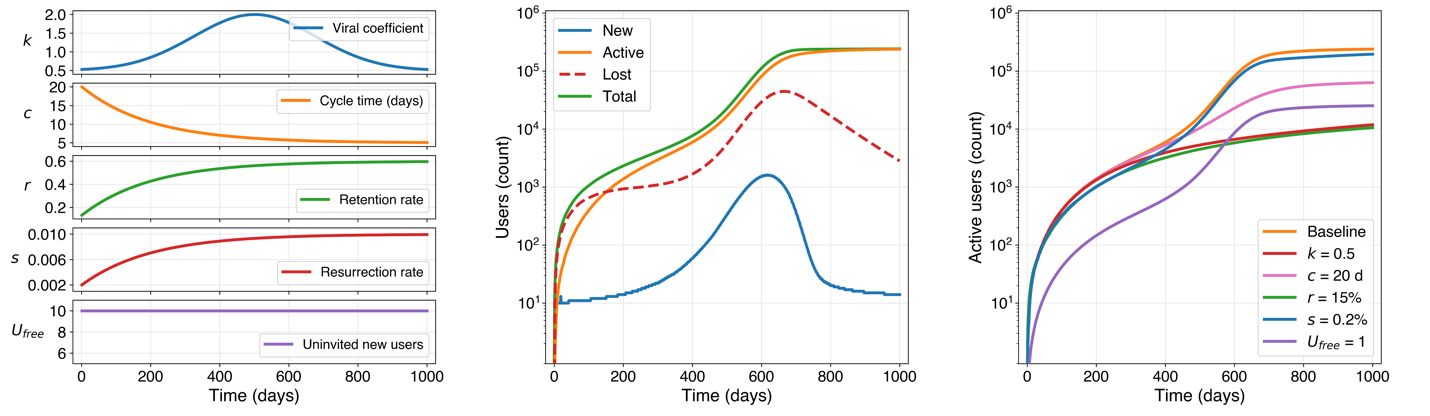
**Total active users** at time 𝑡 is equal to active users from previous interval, plus the previous interval's new users times its retention rate *r*, plus the previous interval's lost users times its resurrection rate *s*:  
  
  
  
This expression makes the simplifying assumption that all churn happens at the new user branch between active and lost categories (exclusively “early churn” as marked in Scheme 2).

**Total lost users** at time 𝑡 is equal to lost users from previous interval plus product of previous interval's new users and churn rate 1 – *r* :



Using these expressions, user growth was modeled for a hypothetical case given by the input parameters shown in the left panel of Figure 5. This scenario aims to replicate the trajectory of a successful company that hits market saturation: cycle time, user retention, and resurrection rates all improve, while viral coefficient experiences an initial increase and subsequent decline to simulate market saturation. This scenario also assumes 10 uninvited new users arrive each day (*U*free = 10).

The result produced from these inputs is shown in the center panel. Growth in active users (orange trace) is initially driven by uninvited new users. Once the viral coefficient exceeds 1, however, the new user count (blue trace) increases dramatically. As *k* drops below 1, new user count again approaches the baseline value and the total user count (green trace) stabilizes. Nonzero resurrection rate pushes some of the lost users (red trace) back into the active category, narrowing the gap between active user count and total registered users.



**Figure 5**. Modeling user growth. Baseline inputs (left) involving viral coefficient reaching a maximum of *k* = 2, cycle time dropping from *c* = 20 to 5 days, retention rate increasing from *r* = 15 to 60%, resurrection rate increasing from s = 0.2 to 1%, and constant arrival of 10 uninvited new users. These inputs were selected to simulate a company making progress against *c*, *r*, and *s*, and *k* in line with the user base reaching market saturation. Corresponding output (center) showing new, active, lost, and total users as a function of time. Comparison of active user count vs. time for the baseline scenario and various sub-optimal cases (right).

The influence of each of the 5 input parameters (*k*, *c*, *r*, *s*, and *U*free) on active user count is shown in the right panel of Figure 5. As compared with the base case, giving up improvements in viral coefficient *k* (red trace) or retention rate *r* (green trace) yields the biggest loss in terminal user count. These curves lack the upward inflection of the others, underscoring the importance of *k* and *r* in turning on viral growth. Retention matters here because of the assumption (Eq. 2) that only retained users, or those who find long-term value in the product will contribute to invites.

On the other hand, giving up improvements to cycle time *c* (pink trace) softens the viral growth bump without eliminating it altogether. Reducing the daily uninvited new users *U*free (violet trace) shrinks the user count compounding during viral growth, resulting in a downward translation of the baseline growth trajectory. Persistently low resurrection rate *s* (blue trace) has only a small detrimental effect on terminal user count.

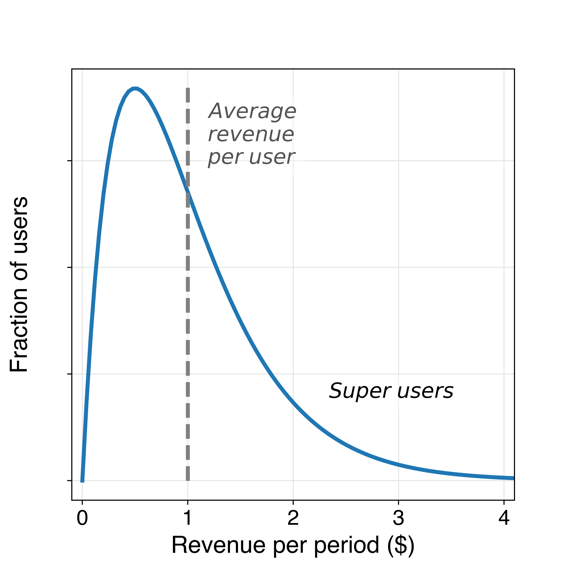
The modeled growth scenarios presented above serve to highlight the role of viral coefficient and retention rate in *turning on* viral growth, how cycle time acts as a *force multiplier* during that period, and how growth outside of the viral period is determined entirely by the arrival of uninvited new users.

***IV. Users and profitability***

The goal of any business is to generate a [profit](https://www.investopedia.com/terms/p/profit.asp), or revenues in excess of costs. This section outlines the connection between a company’s user base and its revenues, costs, and resulting profitability.

*1. Revenue*

A company’s revenue is typically distributed nonuniformly across its user base. For example, businesses generating revenue from advertisements or one-time product sales have a nearly continuous and potentially wide-ranging per-customer revenue spread whose shape can be approximated by a skewed (e.g., gamma or log-normal) distribution (Figure 6). This characteristic, reflecting the fact that some customers contribute a disproportionate share of a company’s revenue, is also [seen in other measures](https://arxiv.org/PS_cache/cond-mat/pdf/0412/0412004v3.pdf) such as household income, book sales, city populations, earthquake sizes, name frequency, and academic citations, among others.



**Figure 6**. Long-tailed distribution of revenues per customer.

Even subscription-based businesses featuring a just a few pricing plans can see a broad set of possible values revenue per user if users are spread across the globe. Netflix, for example, has just 3 pricing tiers but [collects subscription fees](https://www.comparitech.com/blog/vpn-privacy/countries-netflix-cost/) in many currencies and adjusts pricing across single-currency (i.e., Eurozone) regions. Because of this, Netflix users in Switzerland can pay over $22 per month while those in Colombia can pay as little as $5 per month.

Dividing the total revenue over a particular time interval by the active user count (DAUs, MAUs, subscribers, etc.) yields the [average revenue per user](https://www.investopedia.com/terms/a/average-revenue-user-arpu.asp) (ARPU). While it ignores the potentially significant differences in revenue contributions across the user base (Figure 6), ARPU provides a useful link between a company’s user count and its revenue within a given time period.

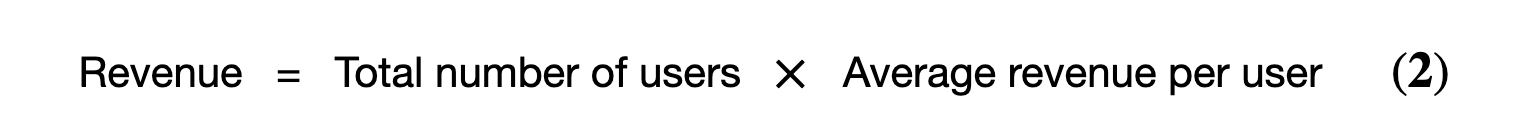
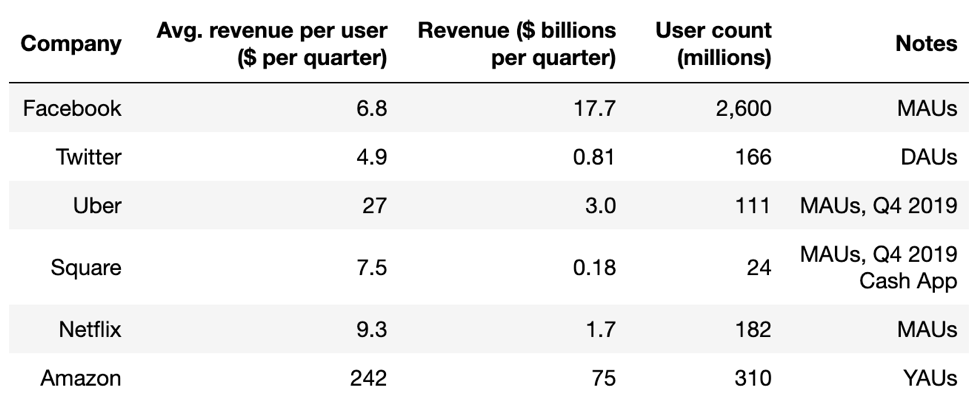


Table 5 shows quarterly ARPU for selected companies in 2020.



**Table 5**. Quarterly revenue per user for selected companies. References: [Facebook](https://s21.q4cdn.com/399680738/files/doc_financials/2020/q1/Q1-2020-FB-Earnings-Presentation.pdf), [Twitter](https://s22.q4cdn.com/826641620/files/doc_financials/2020/q1/Q1-2020-Shareholder-Letter.pdf), [Uber](https://s23.q4cdn.com/407969754/files/doc_financials/2019/sr/InvestorPresentation_2020_Feb13.pdf), [Square](https://s21.q4cdn.com/114365585/files/doc_financials/2019/q4/2019-Q4-Shareholder-Letter-Square.pdf), [Netflix](http://d18rn0p25nwr6d.cloudfront.net/CIK-0001065280/4afe837e-b704-49cd-b9f3-9583bfad10ca.pdf), [Amazon](https://etaileast.wbresearch.com/blog/amazons-engaged-buyers-drive-social-media-revenue).

ARPU can, in turn, be expressed as:



Application of Eq. 3 is straightforward in the case of companies whose revenues come from a series of one-off transactions (e.g., Amazon, Uber, Netflix). For social media, on the other hand, the number of purchases applies to number of ad clicks or impressions, and purchase amount corresponds to the average price fetched from advertisers for cost per click or cost per impression ([CPM or CPC](https://www.criteo.com/insights/whats-difference-cpc-cpm/)) ad pricing.

Increasing revenue is thus a matter of increasing either the user base (Section 1), the average purchase amount, or both.

2. Costs

Variable costs:

* Pre-conversion: marketing expenses – customer acquisition cost
* Post-conversion: customer support, transactions fees, refunds

3. Profitability

* Customer lifetime value
  + Simple definition – assumes constant churn rate,
    - customer lifetime (months) x avg spend / month x gross margin
    - = (1 / churn rate) x avg spend / month [geom. series]
    - Ex. $100 monthly spend x 10% profit margin / 10% monthly churn
    - = ($10 profit / month) x 10 month customer lifetime = $100 CLV

Average cost per user

-fixed costs (spread out across user base)

-variable costs (each user brings)

Monthly recurring revenue (MRR)

Caveat:

-viral coefficient

-cycle time

-caveat: not all new users arrive as a result of contact with an existing user

*Churn*

-retention

-churn rate/probability

-customer lifetime

*Profitability*

-customer lifetime value

-customer acquisition cost

-payback period