Supporting Information

Understanding the group dynamics and success of teams
by Michael Klug and James P. Bagrow

Table of Contents

List of Figures

S1 GitHub popularity

Since roughly 2011 GitHub has become the predominant online platform for developing and hosting open source software and other projects. Competitors such as Sourceforge have been left behind. Figure S1 shows the Google search volume for GitHub and its competitors, as gathered from Google Trends (`https://www.google.com/trends/`).

S2 Statistical analysis of team and success distributions

The empirical distributions of team size $P(M)$, number of teams per person $P(k)$, and success $P(S)$ all appear to be heavy-tailed (Main text Fig. 1). All three empirical distributions are integer-valued, so we use discrete fits. We present complementary cumulative distributions ($P_{>}(X)$) so as to avoid the need to bin the data, which may influence its appearance in a plot.

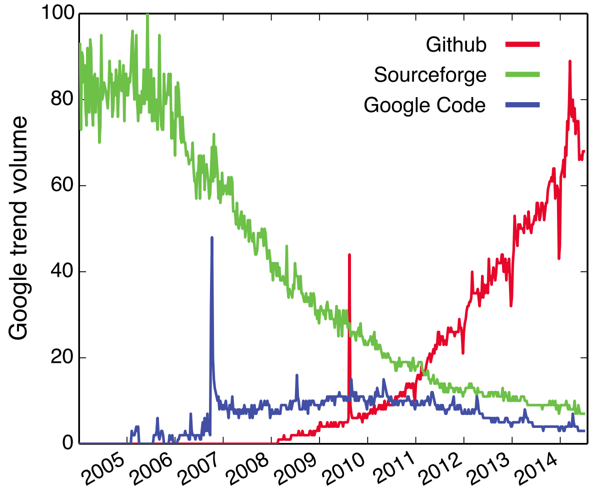

Figure S1: GitHub is becoming the dominant platform for open development and collaboration. Google trends (normalized) search volume over time for searches "GitHub", "Sourceforge", and "Google Code". This trend data serves as a proxy for popularity.

Here we fit to these data three model for heavy-tailed distributions:

$$\text{Power Law (PL): } P(x) \sim x^{-\alpha} \;\forall\; x \geq x_{\min} \tag{S1}$$

$$\text{Log-normal (LN): } P(x) = \frac{1}{x\sqrt{2\pi}\sigma}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \tag{S2}$$

$$\text{Truncated Power Law (TPL): } P(x) \sim x^{-\alpha}\, e^{-x/x_{\text{cut}}} \tag{S3}$$

Note that the TPL is a nested distribution, becoming equivalent to the PL when $x_{\text{cut}} \to \infty$.

The empirical fitting procedure used allows us to estimate both the power law tail exponent α and the power law cutoff $x_{\min}$ [1, 2, 3]. We fit the three models to the data both with the estimated $x_{\min}$ from their procedure and to the entire dataset by forcing $x_{\min} = 1$. Best fit parameters are given in the legends of the respective figures: Figs. S2, S3, and S4.

We compare the distributions using likelihood ratio tests (LRT). This does not confirm one distribution as being correct, but it does allow us to say if one distribution is more likely than another, or if they are statistically indistinguishable.

Team size (Fig. S2) The full $P(M)$ is better explained by either the LN or TPL than the PL. These two distributions are both equally likely (LRT, $p > 0.15$). The tail of the team size distribution ($x_{\min} \geq 7$) is well explained by all three models (LRT, $p > 0.2$). We conclude that M follows a mixed distribution but is likely to have a pure PL tail.

Number of teams per person (Fig. S3) The distribution of team membership, the number of teams k a person belongs to, is better explained by the LN than either the PL or TPL distributions. This holds for the full data (LRT, $p < 10^{-33}$) and for the $x_{\min}$ tail (LRT, $p < 0.04$). We conclude that $P(k)$ is more likely to follow LN than PL or TPL.

Success (Fig. S4) The distribution of success $P(S)$, or the number of GitHub users who have "bookmarked" a team's project. We find, according to the applied fitting procedure, that a TPL model is signif-

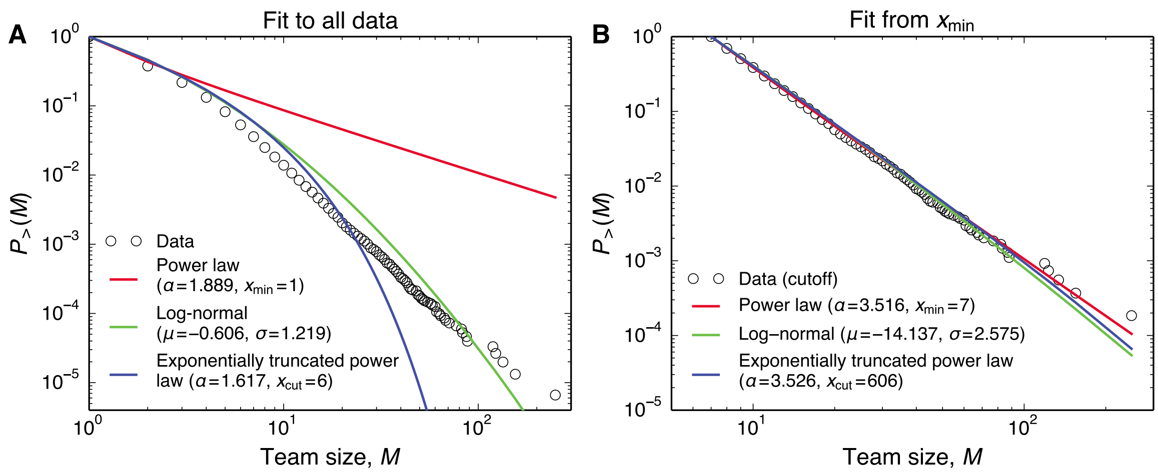

Figure S2: Best fit distributions for team size. (**A**) The full dataset is better explained by either the log-normal or the truncated power law than the pure power law (likelihood ratio test, $p < 10^{-10}$). The truncated power law and the log-normal models are equally likely (likelihood ratio test, $p > 0.15$). (**B**) In the tail of the data all three models are equally likely (Likelihood ratio test, $p > 0.22$) Since the truncated power law has such a large $x_{\rm cut}$ compared to the values of M, we conclude that it is equivalent to the pure power law.

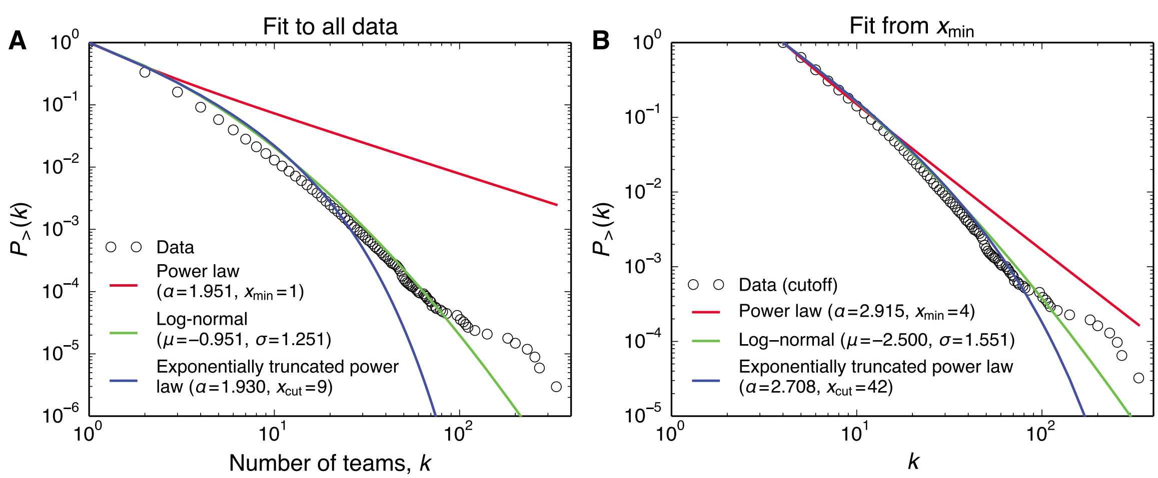

Figure S3: Best fit distributions for number of teams per user. Both the full distribution and the tail ($k > x_{\min}$) are poorly explained by the power law model. The log-normal model is more likely than the truncated power law in both sets of data, although the significance is relatively weak for the tail (likelihood ratio test, $p < 10^{-33}$ in the full data, $p < 0.04$ for the tail data).

icantly more likely than a PL or LN (LRT, $p < 10^{-81}$). (As the tail cutoff was small, $x_{\min} = 2$, we only present fits to the tail.)

S3 The distribution of workload across team members

Figure S5A shows how the total workload (number of pushes) of teams grows with team size M, meaning larger teams are more active in total. This makes sense. The inset shows how the fraction of total work done

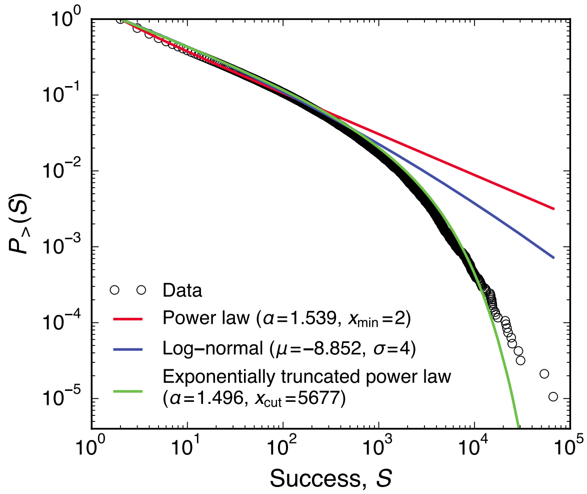

Figure S4: Best fit distributions for success. The truncated power law distribution is significantly more likely than the pure power law or log-normal distribution (likelihood ratio test, $p < 10^{-81}$). As the estimated tail cutoff was small ($x_{\min} = 2$) we only present fits to this region.

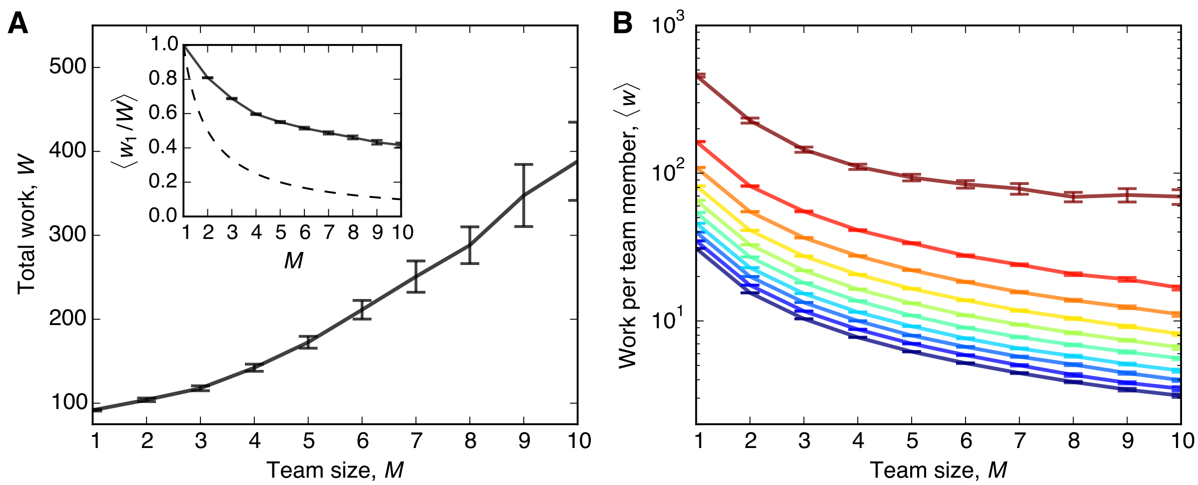

Figure S5: Larger teams are more active but all teams distribute work the same way. (**A**) The total work (total number of pushes) grows with team size, on average. (Inset) The average fraction of work done by the lead decreases with team size, but on average remains well above the lower bound $1/M$. (**B**) The average work per team member decreases as team size grows. The ten curves correspond to the deciles of the total work distribution. The functional form is identical for each decile, except perhaps the highest decile which decays slightly more slowly with M, indicating that the way in which work is distributed over a team is independent of the total activity of that team.

by the lead (w_1/W) decreases on average as team size grows. The lower bound $1/M$ is almost never reached, indicating that workloads remain "front-loaded" for all team sizes.

Figure S5B shows how the average workload per team member *decreases* as teams grow. The ten curves correspond to a decile of the total workload distribution. They all follow the same functional form, except possibly the highest decile which decays slightly less. This indicates that the distribution of workload across team members is independent of total workload; more active teams distribute their workloads across their members in the same manner as less active teams.

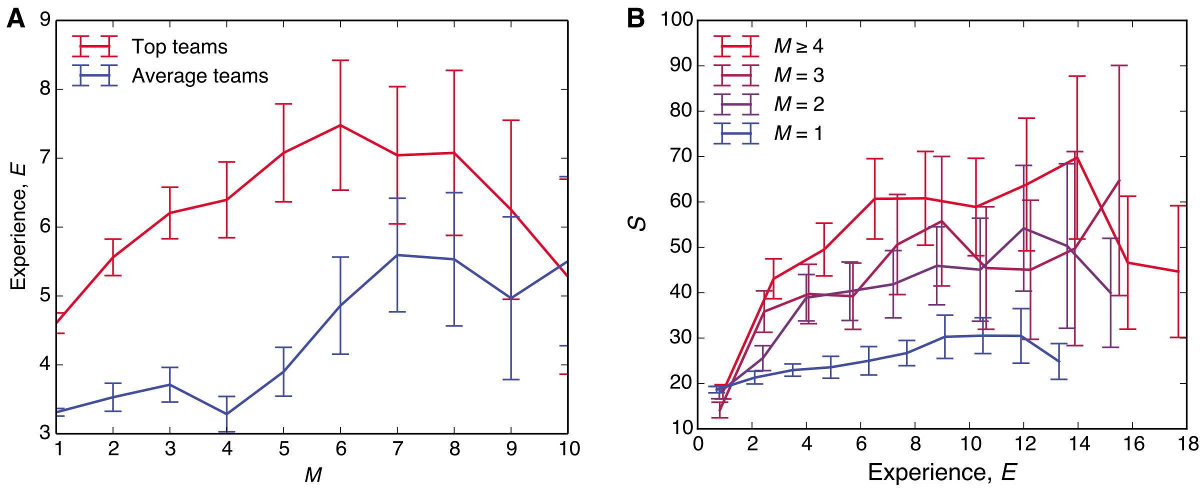

Figure S6: Experience weakly correlates with success. (**A**) Top teams have higher experience than average teams, but only for team sizes $M < 7$. (**B**) As experience grows, success grows on average, but the change in S is quite weak, compared to that of diversity and number of lead members (main text). Linear and nonlinear models indicate that the effects of E are entirely due to the other quantities.

S4 Experience is weakly related to team success

One of the quantities we used to understand a team is their experience E, defined as the average number of other teams that members of the team belong to. We found in the main text that E and S are significantly correlated by themselves, but that E is not significant when including the other variables total team size, effective team size, diversity of experience, and number of lead members. This was shown using both a linear model and a nonlinear technique known as Symbolic Regression.

Here we further underscore these results. In Fig. S6A we find that top teams have significantly higher experience than average teams, but this distinction disappears for $M \geq 7$. Figure S6B shows how E and S relate directly, independent of other quantities. We do see a significant increase in S as E grows, but the change in S is far smaller in magnitude than that of diversity D or number of leads L (shown in the main text). Moreover this effect is confounded with team size, and larger teams may even show a decrease in success as experience grows, although there are insufficient statistics to conclude a trend exists.

S5 Median success vs. team size

In the main text we computed the mean success against various other quantities. Since success is broadly distributed, we removed the top 1% highest success teams from these averages, to mitigate the influence of outliers on the average. Alternatively, we could take the median of success, which is robust against such small outlier effects. We show this in Fig. S7. The median shows the same trend as the mean, although it is highly discretized since most teams have success $S < 20$.

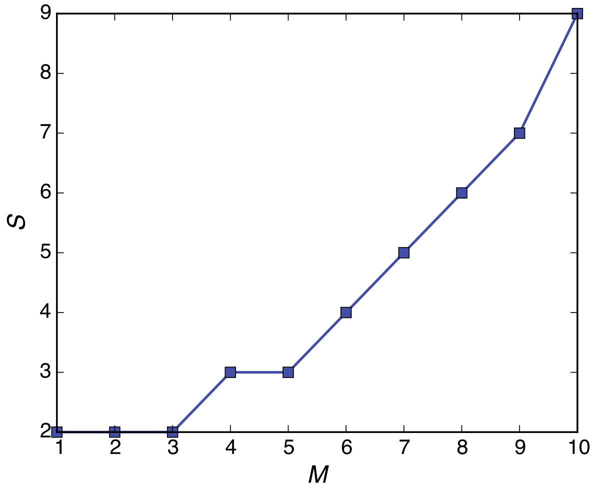

Figure S7: Larger teams have higher median success.

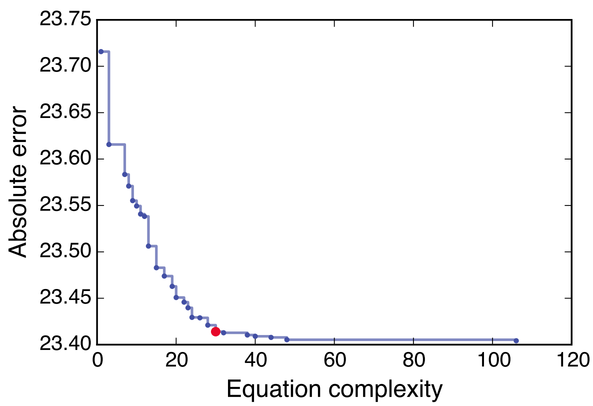

Figure S8: The Pareto Front of optimum equations for given complexities. The point corresponding to the best overall equation is highlighted.

S6 Symbolic regression

In addition to the traditional linear regression model presented in the main text, we also use a new nonlinear modeling technique known as Symbolic Regression (SR) [4, 5]. SR helps us find equations $y = f(x_1, x_2, \ldots)$ without necessarily giving an explicit functional form for f. This makes it a generalization of both linear and non-linear regression techniques.

We used the Eureqa SR package [5] on our team data to find families of functions of the form $S = f(M, m, W, E, D, L)$. Eureqa finds the set of lowest error functions at given complexities, forming a Pareto Front of solutions [4]. Running Eureqa with its default settings on our data, we found 25 functions. A plot of the Pareto front is shown in Fig. S8. The distributions of how many times each variable occurs in each equation is given in Fig. S9. The equations themselves are listed below. (The overall best fit equation is Eq. (S10).)

We find good support for the same conclusions of our linear model. Experience E does not appear in any of the discovered models. Moreover, to determine whether an increase in one variable (such as M or m) causes a positive or negative change in S, we perform a variable sensitivity study (below). This study again supports the results of the linear model. Increasing M, D, and L almost always increases S, whereas

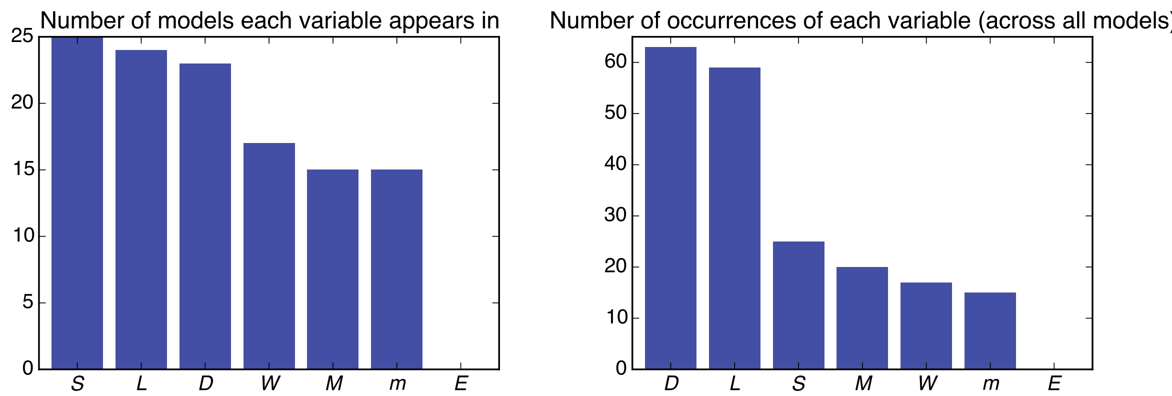

Figure S9: The distribution of variables per equation. There are 25 models total. The left plot counts how many of them contain a particular variable at least once. The right plot counts the total number of occurrences of each variable. Success S must appear exactly once in each equation (on the LHS), so it has value 25. Both D and L appear more than once in equations, with D appearing 2.52 times per equation, on average, and L appearing 2.36 times. Experience E never appears in any model, matching the fact that it was not significant in the linear regression model.

increasing m almost always decreases S. Thus large, diversely experienced, focused teams tend to have the highest success.

S6.1 Variable Sensitivity Study

This section details the complexities and accuracies of the model equations found by Eureqa, as well as the sensitivity of these equations to changes in their variables. This report was generated by the Eureqa 0.99.9 Beta (build 4352) software.

(The best overall equation, according to Eureqa's measure on the tradeoff between complexity and accuracy, is Eq. (S10).)

Explanation of terms:

Sensitivity: The relative impact within this model that a variable has on the target variable.

% Positive: The likelihood that increasing this variable will increase the target variable. If % positive = 70%, then 70% of the time increases in this variable lead to increases in the target variable (but the remaining 30% of the time it either decreases it or has no impact). If % positive = 0%, increases in this variable will not increase the target variable.

Positive Magnitude: When increases in this variable lead to increases in the target variable, this is generally how big the positive impact is.

% Negative: The likelihood that increasing this variable will decrease the target variable. If % negative = 60%, then 60% of the time increases in this variable lead to decreases in the target variable (but the remaining 40% of the time it either increases it or has no impact). If % negative = 0%, increases in this variable will not decrease the target variable.

Negative Magnitude: When increases in this variable lead to decreases in the target variable, this is generally how big the negative impact is.

Details: Given a model equation of the form $z = f(x, y, \ldots)$, the influence metrics of x on z are defined as follows:

Sensitivity: $\overline{\left|\frac{\partial z}{\partial x}\right|} \cdot \frac{\sigma(x)}{\sigma(z)}$, evaluated at all input data points;

% Positive: The percent of data points where $\frac{\partial z}{\partial x} > 0$;

% Negative: The percent of data points where $\frac{\partial z}{\partial x} < 0$;

Positive magnitude: $\overline{\left|\frac{\partial z}{\partial x}\right|} \cdot \frac{\sigma(x)}{\sigma(z)}$, at all points where $\frac{\partial z}{\partial x} > 0$;

Negative magnitude: $\overline{\left|\frac{\partial z}{\partial x}\right|} \cdot \frac{\sigma(x)}{\sigma(z)}$, at all points where $\frac{\partial z}{\partial x} < 0$;

where $\frac{\partial z}{\partial x}$ is the partial derivative of z with respect to x, $\sigma(x)$ is the standard deviation of x in the input data, $\sigma(z)$ is the standard deviation of z, $|x|$ denotes the absolute value of x, and $\overline{x}$ denotes the mean of x.

$$S = 0.00086313L^{2.963D}W - 0.090191L^{2.963D} + \frac{1}{m}\left(1.1705^{M-0.56145}D^{3.5035D}L^{3.4195^{D}} + M\right) \quad \text{(S4)}$$

Complexity = 106

Absolute Error = 23.404200

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.50409	100%	0.50409	0%	0
m	0.38949	0%	0	100%	0.38949
M	0.34811	100%	0.34811	0%	0
D	0.33048	96%	0.34244	4%	0.010864
W	0.056825	100%	0.056825	0%	0

$$S = 0.00086858L^{2.9785D}W - 0.09076L^{2.9785D} + \frac{1}{m}\left(1.1848^{M-0.57673}D^{3.2116D}L^{3.2856D} + M\right) \quad \text{(S5)}$$

Complexity = 48

Absolute Error = 23.405300

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.47639	100%	0.47639	0%	0
m	0.38147	0%	0	100%	0.38147
M	0.34609	100%	0.34609	0%	0
D	0.31481	97%	0.32551	3%	0.010681
W	0.056258	100%	0.056258	0%	0

$$S = 0.00086817L^{2.9971D}W - 0.090717L^{2.9971D} + \frac{1}{m}\left(D^{3.5411D}L^{3.303D}M^{0.5} + M\right) \quad \text{(S6)}$$

Complexity = 44

Absolute Error = 23.407800

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.51542	100%	0.51542	0%	0
M	0.46004	100%	0.46004	0%	0
m	0.42114	0%	0	100%	0.42114
D	0.36405	97%	0.37649	3%	0.011786
W	0.063722	100%	0.063722	0%	0

$$S = 0.00083731L^{3.0424D}W - 0.087492L^{3.0424D} + \frac{1}{m}\left(1.1725^{M-0.62723}DL^{2.9285D} + M\right) \tag{S7}$$

Complexity = 40

Absolute Error = 23.409000

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.52306	100%	0.52306	0%	0
m	0.46692	0%	0	100%	0.46692
M	0.4168	100%	0.4168	0%	0
D	0.2472	100%	0.2472	0%	0
W	0.068029	100%	0.068029	0%	0

$$S = 0.00086313L^{3.0311D}W - 0.090191L^{3.0311D} + \frac{1}{m}\left(1.0822^{M}DL^{3.1701D} + M\right) \tag{S8}$$

Complexity = 38

Absolute Error = 23.410400

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.61212	100%	0.61227	0%	0.00084022
m	0.50192	0%	0	100%	0.50192
M	0.39633	100%	0.39633	0%	0
D	0.26478	100%	0.26478	0%	0
W	0.074642	100%	0.074642	0%	0

$$S = 0.00086817L^{3.0339D}W - 0.090717L^{3.0339D} + \frac{1}{m}\left(1.0606DL^{3.4478D} + M\right) \tag{S9}$$

Complexity = 32

Absolute Error = 23.412800

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.67506	100%	0.67575	0%	0.0012525
m	0.51434	0%	0	100%	0.51434
M	0.36037	100%	0.36037	0%	0
D	0.27647	100%	0.27647	0%	0
W	0.077665	100%	0.077665	0%	0

$$S = 0.00086817L^{3.0321D}W - 0.090717L^{3.0321D} + \frac{1}{m}\left(DL^{3.5006D} + M\right) \tag{S10}$$

Complexity = 30

Absolute Error = 23.414000

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.64634	100%	0.64711	0%	0.0012379
m	0.49821	0%	0	100%	0.49821
M	0.35758	100%	0.35758	0%	0
D	0.27078	100%	0.27078	0%	0
W	0.076992	100%	0.076992	0%	0

$$S = 0.00083712L^{3.0861D}W - 0.087472L^{3.0861D} + \frac{1}{m}\left(L^{3.0863D} + M\right) \tag{S11}$$

Complexity = 28

Absolute Error = 23.421000

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.70447	100%	0.70447	0%	0
m	0.60966	0%	0	100%	0.60966
M	0.43345	100%	0.43345	0%	0
D	0.18037	18%	0.97783	0%	0
W	0.092313	100%	0.092313	0%	0

$$S = 0.00094168L^{3.0422D}W - 0.098398L^{3.0422D} + \frac{1}{m}\left(1.1008DL^{2} + M\right) \tag{S12}$$

Complexity = 26

Absolute Error = 23.428900

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
m	0.80679	0%	0	100%	0.80679
L	0.66859	100%	0.66861	0%	0.15262
M	0.59982	100%	0.59982	0%	0
D	0.19448	100%	0.19432	0%	0.28059
W	0.14074	100%	0.14074	0%	0

$$S = \frac{2.0512D}{m}LM + 0.0007675L^{3.1623D}W - 0.080198L^{3.1623D} \tag{S13}$$

Complexity = 24

Absolute Error = 23.429400

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
M	0.92419	100%	0.92419	0%	0
m	0.6244	0%	0	100%	0.6244
L	0.52968	100%	0.52968	0%	0.17883
D	0.29889	100%	0.29886	0%	0.53101
W	0.089784	100%	0.089784	0%	0

$$S = 2.0569DL + 0.00072728L^{3.2045D}W - 0.075995L^{3.2045D} + M - m \tag{S14}$$

Complexity = 23

Absolute Error = 23.439700

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
M	0.70532	100%	0.70532	0%	0
L	0.61753	100%	0.61757	0%	0.18733
m	0.40931	0%	0	100%	0.40931
D	0.30963	100%	0.30953	0%	0.52775
W	0.11459	100%	0.11459	0%	0

$$S = 0.00073349L^{3.2053D}W - 0.076643L^{3.2053D} + \frac{1}{m}(D + LM) \tag{S15}$$

Complexity = 22

Absolute Error = 23.445800

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
m	0.68988	0%	0	100%	0.68988
M	0.60299	100%	0.60299	0%	0
L	0.37671	100%	0.37678	0%	0.11478
D	0.11739	95%	0.11806	5%	0.10395
W	0.10301	100%	0.10301	0%	0

$$S = L + \frac{1}{m}\left(L + 0.0009206L^{4.0335D}W - 0.096195L^{4.0335D}\right) \tag{S16}$$

Complexity = 20

Absolute Error = 23.450800

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.44438	100%	0.44482	0%	0.17098
m	0.24286	0%	0.084319	100%	0.24312
W	0.089414	100%	0.089414	0%	0
D	0.056473	7%	0.67228	11%	0.078402

$$S = DL + L + 0.00091119L^{3.0604D}W - 0.095213L^{3.0604D} \tag{S17}$$

Complexity = 19

Absolute Error = 23.462800

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.70835	100%	0.7084	0%	0.16293
D	0.18624	100%	0.18588	0%	0.27669
W	0.14558	100%	0.14558	0%	0

$$S = D + L + 0.0010251L^{3.0304D}W - 0.10712L^{3.0304D} \tag{S18}$$

Complexity = 17

Absolute Error = 23.473900

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.42938	100%	0.42984	0%	0.13934
D	0.18274	98%	0.18181	2%	0.21919
W	0.17992	100%	0.17992	0%	0

$$S = 0.0014928LW - 0.15599L + 0.79229e^{DL} \tag{S19}$$

Complexity = 15

Absolute Error = 23.482900

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.22823	100%	0.22823	0%	0.0011273
D	0.16673	100%	0.16673	0%	0
W	0.038473	100%	0.038473	0%	0

$$S = 0.0018278W + 0.80776e^{DL} - 0.19099 \tag{S20}$$

Complexity = 13

Absolute Error = 23.506200

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.2308	100%	0.2308	0%	0
D	0.16862	100%	0.16862	0%	0
W	0.036139	100%	0.036139	0%	0

$$S = e^{DL} - \frac{0.71828}{M} \tag{S21}$$

Complexity = 12

Absolute Error = 23.538200

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.23073	100%	0.23073	0%	0
D	0.16856	100%	0.16856	0%	0
M	0.081472	100%	0.081472	0%	0

$$S = -D + e^{DL} + 0.28172 \tag{S22}$$

Complexity = 11

Absolute Error = 23.540700

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.23133	100%	0.23133	0%	0
D	0.14482	100%	0.14482	0%	0

$$S = \frac{2.0D}{m} LM \tag{S23}$$

Complexity = 10

Absolute Error = 23.549400

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
M	1.2657	100%	1.2657	0%	0
m	0.85515	0%	0	100%	0.85515
L	0.71957	100%	0.71957	0%	0
D	0.3732	100%	0.3732	0%	0

$$S = L + (DL)! \tag{S24}$$

Complexity = 9

Absolute Error = 23.555200

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
D	0.033094	95%	0.034694	5%	0.00065118
L	0.032843	100%	0.032843	0%	0

$$S = D + \frac{LM}{m} \tag{S25}$$

Complexity = 8

Absolute Error = 23.570900

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
M	0.86537	100%	0.86537	0%	0
m	0.61724	0%	0	100%	0.61724
L	0.53289	100%	0.53289	0%	0
D	0.13195	100%	0.13195	0%	0

$$S = e^{DL} \tag{S26}$$

Complexity = 7

Absolute Error = 23.583400

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	0.2315	100%	0.2315	0%	0
D	0.16912	100%	0.16912	0%	0

$$S = 2.0L \tag{S27}$$

Complexity = 3

Absolute Error = 23.615700

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
L	1	100%	1	0%	0

$$S = 2.0 \tag{S28}$$

Complexity = 1

Absolute Error = 23.715800

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
This model contains no variable references.					

References

[1] A. Clauset, C. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, 2009.

[2] A. Klaus, S. Yu, and D. Plenz. Statistical analyses support power law distributions found in neuronal avalanches. *PLoS ONE*, 6(5):e19779, 05 2011.

[3] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777, 2014.

[4] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[5] M. Schmidt and H. Lipson. Eureqa (version 0.99 beta)[software], 2014.