Complete asymptotic type-token relationship for growing complex systems with inverse power-law count rankings

Pablo Rosillo-Rodes,[1, 2, *] Laurent Hébert-Dufresne,[3, 4, †] and Peter Sheridan Dodds[2, 3, 4, ‡]

[1] *Institute for Cross-Disciplinary Physics and Complex Systems IFISC (UIB-CSIC), Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain*
[2] *Computational Story Lab, University of Vermont, Burlington, VT 05405, US*
[3] *Department of Computer Science, Vermont Complex Systems Institute, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Center, University of Vermont, Burlington, VT 05405, US*
[4] *Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, US*

(Dated: November 5, 2025)

The growth dynamics of complex systems often exhibit statistical regularities involving power-law relationships. For real finite complex systems formed by countable tokens (animals, words) as instances of distinct types (species, dictionary entries), an inverse power-law scaling $S \sim r^{-\alpha}$ between type count S and type rank r, widely known as Zipf's law, is widely observed to varying degrees of fidelity. A secondary, summary relationship is Heaps' law, which states that the number of types scales sublinearly with the total number of observed tokens present in a growing system. Here, we propose an idealized model of a growing system that (1) deterministically produces arbitrary inverse power-law count rankings for types, and (2) allows us to determine the exact asymptotics of the type-token relationship. Our argument improves upon and remedies earlier work. We obtain a unified asymptotic expression for all values of α, which corrects the special cases of $\alpha = 1$ and $\alpha \gg 1$. Our approach relies solely on the form of count rankings, avoids unnecessary approximations, and does not involve any stochastic mechanisms or sampling processes. We thereby demonstrate that a general type-token relationship arises solely as a consequence of Zipf's law.

INTRODUCTION

Universal statistical regularities play a central role in the study of complex systems. Prominent examples are power-law relationships, such as Zipf's law for word frequencies [1, 2], observed across human [3] and animal [4] communication. For language, Zipf's law states that the count of a word in a text decreases with its rank r as $S_{r,\alpha} \sim r^{-\alpha}$, originally with $\alpha \simeq 1$. While arguments about the empirical validity and mechanistic origins of Zipf's law have remained contentious for the better part of a century [5–10], count-rankings adhering approximately to an inverse power-law have been observed across systems of all kinds including ecological systems [5, 11], complex networks [12, 13], and socioeconomic systems [14–17].

Another well-known scaling regularity is Heaps' law [18–20], which describes how the vocabulary count N of a text grows with its total length t as $N_{t,\beta} \sim t^{\beta}$, with $0 < \beta < 1$. Like Zipf's law, type-token scaling has been found to generalize beyond language, and has been reported to hold in, for example, chemoinformatics [21], computer code [22], and urban systems [23]. The type-token relationship is a summary one, as it carries far less information than count ranking which records not just type ranks but the names of the types themselves.

The fact that Zipf's and Heaps' laws are empirically observed to hold simultaneously has motivated interest in whether or not a fundamental relationship between them exists in finite, real systems. Previous works have developed a range of frameworks to relate both scaling laws, for example, by empirical observation [24], proposing specific language models [25, 26], growth dynamics [27], sampling mechanisms [28, 29], or assuming in advance their simultaneous validity [30]. Lü *et al.* [31], whose work we improve upon, found that the scaling relation between the two exponents in the infinite system count limit is given by

$$\beta = \begin{cases} 1 & \text{for } \alpha \leq 1, \\ 1/\alpha & \text{for } \alpha > 1. \end{cases} \tag{1}$$

In this work, we derive a complete asymptotic type–token relationship for idealized growing finite systems, valid for any value of the scaling exponent $0 \leq \alpha < \infty$. Our formulation corrects earlier approximations and returns the scaling form of Heaps' law in specific limits. Our approach is independent of the underlying mechanisms of the system, provides an excellent fit to our growing system model, and offers further insight into the growth dynamics of real systems.

MODEL OF AN IDEALIZED GROWING SYSTEM

We consider an idealized growing system comprising countable tokens as instances of types. We set aside physical mechanisms and consider a system emerging in time as follows. The system grows in discrete time so that there are t total tokens at time $t = 1, 2, 3, \ldots$, and such that the count of the rth type is

$$S_{r,t,\alpha} = \left\lfloor \frac{1}{2} + G_{t,\alpha} r^{-\alpha} \right\rfloor, \tag{2}$$

where $\alpha > 0$, $\lfloor \cdot \rfloor$ is the floor operator, converted to a round-to-the-nearest-integer operator by the addition of 1/2, and $G_{t,\alpha}$ is a growth factor which we define below. We indicate count by S which stands for the more general conception of size. For $\alpha = 0$, $S_{r,t,0} = 1$ for $1 \le r \le t$ and 0 otherwise. By construction, types remain ordered in count ranking according to their arrival time. When the rth type's count $S_{r,t,\alpha}$ first rounds up to 1 (i.e., reaches 1/2), we may consider the type as being created or uncovered. At any time t, there will be a finite number $N_{t,\alpha}$ of types with one or more tokens in the system, $N_{t,\alpha}$ being also the rank of the most recently arrived type. Because the system is a realized entity, we call $S_{r,t,\alpha}$ a count-ranking rather than a count-rank distribution, which mis-implies, for our model, a sampling process.

The growth factor $G_{t,\alpha} \ge 0$ "grows" the system in a step-wise fashion so that

$$t = \sum_{r=1}^{\infty} \left\lfloor \frac{1}{2} + G_{t,\alpha} r^{-\alpha} \right\rfloor. \tag{3}$$

Equation (3) means that, for a given value of t, there will be an increasingly small range of values of $G_{t,\alpha}$ for which t tokens are present in the system. Also, not all individual token counts t are achievable, because two types can reach a half integer simultaneously.

Now, while our idealized system can be easily grown computationally, the floor operator prevents ready analysis. We approximate the count-ranking of our model as

$$S_{r,t,\alpha} = \begin{cases} S_{1,t,\alpha} r^{-\alpha} & \text{for} \quad 1 \le r \le N_{t,\alpha}, \\ 0 & \text{for} \quad r > N_{t,\alpha}. \end{cases} \tag{4}$$

Counts are now fractional, and we determine $S_{1,t,\alpha}$ below. We note that in some limits, the approximation may not allow us to capture the idealized system, particularly as $\alpha \to 0$, because the continuous form smooths over the discrete, step-wise growth that dominates when the count hierarchy flattens, causing the approximate scaling to fail to represent the actual ranked increments. Nevertheless, we will be able to satisfactorily connect count-rank scaling to type-token scaling.

To maintain the connection to our idealized base model, we enforce two conditions. First, when a type first appears it has a count of 1. Setting rank r to $N_{t,\alpha}$ in Eq. (4), we have

$$S_{N_{t,\alpha},t,\alpha} = 1 = S_{1,t,\alpha} N_{t,\alpha}^{-\alpha}, \tag{5}$$

which means

$$S_{1,t,\alpha} \sim N_{t,\alpha}^{\alpha}. \tag{6}$$

Second, we must have the sum of fractional counts equal to time t, so

$$t = \sum_{r=1}^{N} S_{r,t,\alpha} \sim N_{t,\alpha}^{\alpha}, H_{N_{t,\alpha}}^{(\alpha)} \tag{7}$$

where

$$H_n^{(a)} = \sum_{k=1}^{n} k^{-a} \tag{8}$$

is the nth generalized harmonic number of order a. Eq. (7) implicitly gives us what we call the type-token relationship, i.e., how the number of distinct types $N_{t,\alpha}$ grows with the overall number tokens, t. We may approximate the harmonic sum in Eq. (7) by its integral form,

$$H_{N_{t,\alpha}}^{(\alpha)} \sim \int_{z=1}^{N_{t,\alpha}} z^{-\alpha} \mathrm{d}z, \tag{9}$$

so we would arrive to

$$t \sim \frac{N_{t,\alpha}^{\alpha}}{1-\alpha} \left[N_{t,\alpha}^{1-\alpha} - 1 \right]. \tag{10}$$

Equation (10) coincides with the result obtained by Lü *et al.* in Ref. [31], where the authors derive the connection between the number of distinct types and the total number of tokens by transforming their count-ranking into a size-frequency distribution (which has exponent $\gamma = 1 + 1/\alpha$) and approximating $H_{N_{t,\alpha}}^{(\alpha)}$ by Eq. (9). Here, we have achieved a direct and compact derivation which proceeds exclusively from the count-rank representation.

However, Eq. (10) is incorrect for large α, and the error is in the integral approximation of Eq. (9) (which we explain below). We return to Eq. (7), and instead use the Euler-Maclaurin expansion [32] of $H_{N_{t,\alpha}}^{(\alpha)}$ for $\alpha \ne 1$, i.e.,

$$\begin{aligned} H_{N_{t,\alpha}}^{(\alpha)} = \frac{N_{t,\alpha}^{1-\alpha}}{1-\alpha} + \zeta(\alpha) + \frac{1}{2} N_{t,\alpha}^{-\alpha} \\ - \frac{\alpha}{12} N_{t,\alpha}^{-\alpha-1} + \mathcal{O}\left(N_{t,\alpha}^{-\alpha-2}\right). \end{aligned} \tag{11}$$

We now have

$$t \simeq \frac{N_{t,\alpha}}{1-\alpha} + N_{t,\alpha}^{\alpha} \zeta(\alpha) + \frac{1}{2} - \frac{\alpha}{12 N_{t,\alpha}} + \mathcal{O}\left(\frac{1}{N_{t,\alpha}^2}\right), \tag{12}$$

which, to leading order, yields

$$t \sim \frac{1}{1-\alpha} N_{t,\alpha} + \zeta(\alpha) N_{t,\alpha}^{\alpha}, \tag{13}$$

where $\zeta(\alpha)$ is the Riemann zeta function for $\alpha > 1$ and its analytical continuation for $\alpha < 1$.

In panels Fig. 1 a)–f), we show how Eq. (13) fits the behavior of Eq. (7) for five example values of α. As we show in the following, we can invert the limiting form of Eq. (13) for certain values and ranges of α and compute $N_{t,\alpha}(t)$ as a function of t.

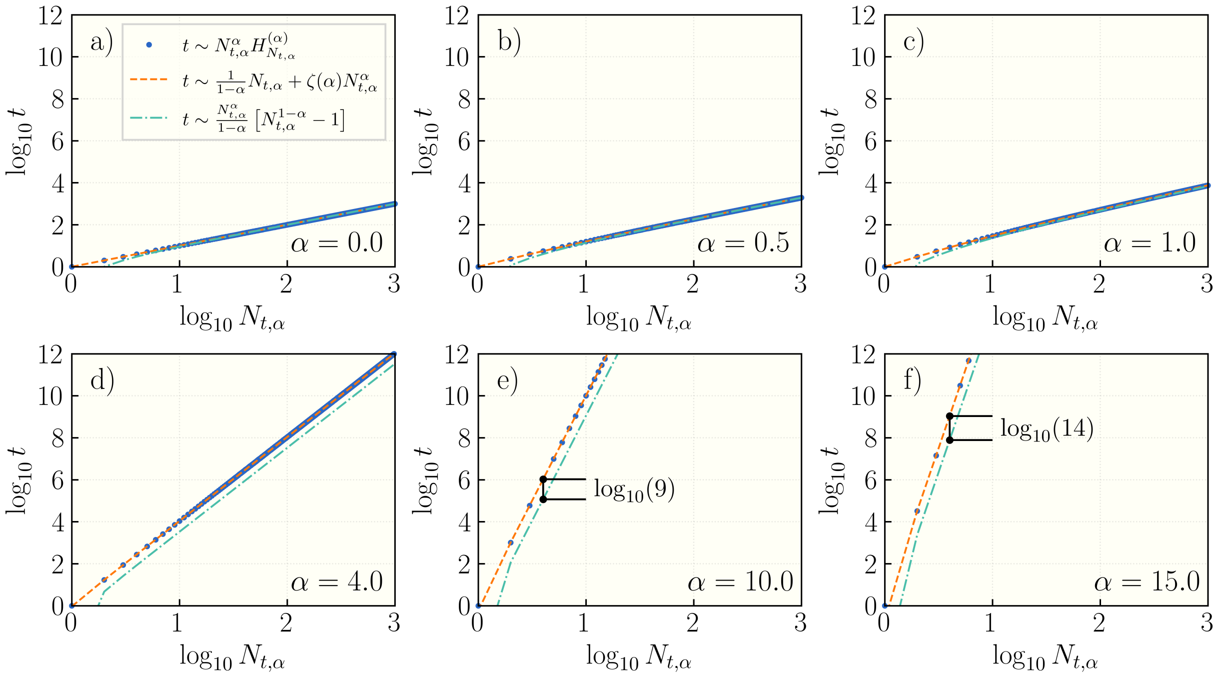

FIG. 1. Comparison between data discretely generated using the harmonic sum in Eq. (7) and the results yielded both by the expansion in Eq. (13) and by Eq. (10) for different values of α. In c), Eq. (20) for $\alpha = 1$ is used to solve the divergence. In e) and f), we show the deviation of Eq. (10) from the data in the $\alpha \gg 1$ regime.

LIMITING BEHAVIOR

For large t and $N_{t,\alpha}$, we can express Eq. (13) in simpler forms for the three regimes of $\alpha \ll 1$, $\alpha \simeq 1$, and $\alpha \gg 1$.

First, for $\alpha \ll 1$, Eq. (12) yields

$$t \sim \frac{1}{1-\alpha} N_{t,\alpha}, \tag{14}$$

because $N^{1}_{t,\alpha} \gg N^{\alpha}_{t,\alpha}$ and $\lim_{\alpha \to 0} \zeta(\alpha) = -1/2$.

Second, for $\alpha \to 1$, we use the approximations

$$N^{\alpha}_{t,\alpha} \to N_{t,\alpha} \left[1 - (1-\alpha) \ln N_{t,\alpha} + \mathcal{O}\left((1-\alpha)^2\right)\right] \tag{15}$$

and

$$(1-\alpha)\zeta(\alpha) \to -1 + \gamma_0(1-\alpha) + \mathcal{O}\left((1-\alpha)^2\right) \tag{16}$$

where $\gamma_0 \simeq 0.577$ is the Euler-Mascheroni constant. Substituting Eqs. (15) and (16) into Eq. (13), we have for $\alpha \to 1$ that

$$t \sim N_{t,\alpha} \left(\ln N_{t,\alpha} + \gamma_0\right). \tag{17}$$

By next using the leading term of the asymptotic form of the Lambert W function, we obtain

$$N_{t,\alpha} \sim \frac{t}{\ln t + \gamma_0 - \ln\left(\ln t + \gamma_0\right)}. \tag{18}$$

Finally, for $\alpha \gg 1$, the exponential term in Eq. (13) dominates and because $\zeta(\alpha) \to 1$, the limiting behavior is given by

$$t \sim N^{\alpha}_{t,\alpha}. \tag{19}$$

We observe that in the regime in which $\alpha \gg 1$, Eq. (10) incorrectly gives $t \to (\alpha-1)^{-1} N^{\alpha}_{t,\alpha}$, which is off from the actual behavior of Eq. (7) by a factor $(\alpha-1)^{-1}$. The reason for the error is that the first term of $H^{(\alpha)}_{N_{t,\alpha}}$ is 1, independent of α, but the steep decay for high α means that the integral approximation of Eq. (10) estimates it as $1/(\alpha-1)$. We indicate the error in panels e) and f) of Fig. 1 for $\alpha = 10$ and $\alpha = 15$, respectively. Note that we maintain $(\alpha-1)^{-1}$ instead of α^{-1} to match our simulations for these values of α.

The above derived limiting forms of Eq. (13) for large t and $N_{t,\alpha}$, can be summarized as

$$t \sim \begin{cases} \frac{1}{1-\alpha} N_{t,\alpha} & \text{for } \alpha \ll 1, \\ N_{t,\alpha}\left(\ln N_{t,\alpha} + \gamma_0\right) & \text{for } \alpha = 1, \\ N^{\alpha}_{t,\alpha} & \text{for } \alpha \gg 1, \end{cases} \tag{20}$$

and

$$N_{t,\alpha} \to \begin{cases} (1-\alpha)t & \text{for } \alpha \ll 1, \\ t/\left[\ln t + \gamma_0 - \ln\left(\ln t + \gamma_0\right)\right] & \text{for } \alpha = 1, \\ t^{1/\alpha} & \text{for } \alpha \gg 1. \end{cases} \tag{21}$$

While Eq. (13) fits the behavior correctly for $\alpha \neq 1$, the applicability of the simpler expressions for the limits in Eqs. (20) and (21) depends on the range of α, as we demonstrate in Fig. 2.

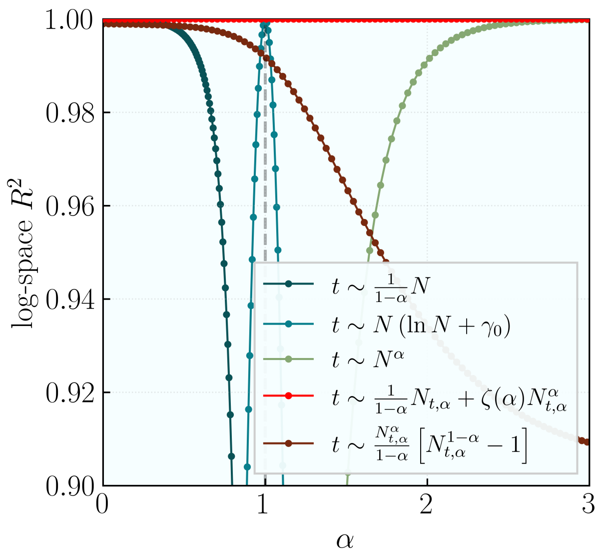

FIG. 2. Goodness of the fit of the expansion in Eq. (13) and the approximations in Eq. (20). The data used for the fit corresponds to 10^3 points for $N_{t,\alpha} \in [1, 10^3]$, computed using the sum in Eq. (7).

DISCUSSION

In this work, we have presented a simple derivation of the asymptotic behavior of the type–token relationship for growing systems with inverse power-law count rankings. Our expression for the number of tokens t (equivalently system size) as a function of the number of distinct types $N_{t,\alpha}$ in Eq. (13) fully captures large system behavior for all $\alpha \geq 0$. Our results extend and correct those of Lü *et al.* [31], achieving a more accurate approximation while only requiring count-ranking. Importantly, because our results follow from an idealized growing model, Eq. (13) is independent of the underlying mechanism that drives real system growth.

As can be visually appreciated in Fig. 1 a)–f), a linear approximation of the type–token relation in log–log space provides an accurate graphical description for any value of α, provided that the system size t is sufficiently large. This observation, as discussed in Ref. [31], clarifies why Heaps' law so frequently appears in empirical studies.

Our results demonstrate that Heaps' law is not an independent phenomenon, or one dependent on stochastic growth, but is rather an emergent statistical relationship of growing systems with inverse power-law count-rankings.

ACKNOWLEDGMENTS

P.R. acknowledges support by the Spanish State Research Agency (MICIU/AEI/10.13039/501100011033) and FEDER (UE) under the María de Maeztu project CEX2021-001164-M and the projects PID2021-122256NB-C21 and PID2024-157493NB-C21, during a research stay in the Computational Story Lab at the University of Vermont. The authors are grateful for National Science Foundation Award #2242829 (Science of Online Corpora, Knowledge, and Stories) and award #2419733, foundational support from MassMutual, and an anonymous philanthropic gift.

* prosillo@ifisc.uib-csic.es
† laurent.hebert-dufresne@uvm.edu
‡ peter.dodds@uvm.edu

[1] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort* (Addison-Wesley, Cambridge, MA, 1949).
[2] M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, Contemporary Physics **46**, 323 (2005).
[3] E. G. Altmann and M. Gerlach, Statistical Laws in Linguistics, in *Creativity and Universality in Language*, edited by M. Degli Esposti, E. G. Altmann, and F. Pachet (Springer International Publishing, Cham, 2016) pp. 7–26.
[4] I. Arnon, S. Kirby, J. A. Allen, C. Garrigue, E. L. Carroll, and E. C. Garland, Whale song shows language-like statistical structure, Science **387**, 649 (2025).
[5] G. U. Yule, A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S., Phil. Trans. B **213**, 21 (1925).
[6] B. B. Mandelbrot, An informational theory of the statistical structure of languages, in *Communication Theory*, edited by W. Jackson (Butterworth, Woburn, MA, 1953) pp. 486–502.
[7] H. A. Simon, On a class of skew distribution functions, Biometrika **42**, 425 (1955).
[8] T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh, Empirical tests of Zipf's law mechanism in open source Linux distribution, Phys. Rev. Lett. **101**, 218701 (2008).
[9] R. Ferrer-i-Cancho and B. Elvevåg, Random texts do not exhibit the real Zipf's law-like rank distribution, PLoS ONE **5**, e9411 (2010).
[10] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, Text mixing shapes the anatomy of rank-frequency distributions, Physical Review E **91**, 052811 (2015).
[11] J. Camacho and R. V. Solé, Scaling in ecological size spectra, Europhysics Letters **55**, 774 (2001).
[12] S. Redner, How popular is your paper? An empirical study of the citation distribution, The European Physical Journal B **4**, 131 (1998).
[13] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).
[14] M. A. Serrano and M. Boguñá, Topology of the world trade web, Phys. Rev. E **68**, 015101 (2003).
[15] C. M. Pinto, A. M. Lopes, and J. T. Machado, A review of power laws in real life phenomena, Communications in Nonlinear Science and Numerical Simulation **17**, 3558

(2012).
[16] X. Gabaix, Power laws in economics: An introduction, Journal of Economic Perspectives **30**, 185 (2016).
[17] A. Gaillard, P. Wangner, C. Hellwig, and N. Werquin, Consumption, wealth, and income inequality: A tale of tails, Available at SSRN 4636704 (2023).
[18] G. Herdan, *Type-token Mathematics*, Janua linguarum: Series maior, Vol. 4 (Mouton, The Hague, 1960) p. 448, series: Studia memoriae Nicolai van Wijk dedicata. Series maior, 4.
[19] H. Heaps, *Information Retrieval, Computational and Theoretical Aspects*, Library and information science (Academic Press, 1978).
[20] A. Chacoma and D. H. Zanette, Heaps' law and Heaps' functions in tagged texts: evidences of their linguistic relevance, Royal Society Open Science **7**, 200008 (2020).
[21] R. W. Benz, S. J. Swamidass, and P. Baldi, Discovery of power-laws in chemical space, Journal of Chemical Information and Modeling **48**, 1138 (2008).
[22] H. Zhang, Discovering power laws in computer programs, Information Processing & Management **45**, 477 (2009).
[23] F. Simini and C. James, Testing Heaps' law for cities using administrative and gridded population data sets, EPJ Data Science **8**, 1 (2019).
[24] R. Baeza–Yates and G. Navarro, Block addressing indices for approximate text retrieval, Journal of the American Society for Information Science **51**, 69 (2000).
[25] D. Zanette and M. Montemurro, Dynamics of text generation with realistic zipf's distribution, Journal of Quantitative Linguistics **12**, 29 (2005).
[26] M. Á. Serrano, A. Flammini, and F. Menczer, Modeling statistical properties of written text, PLOS ONE **4**, e5372 (2009).
[27] L. Hébert-Dufresne, A. Allard, J.-G. Young, and L. J. Dubé, Constrained growth of complex scale-independent systems, Phys. Rev. E **93**, 032304 (2016).
[28] D. van Leijenhorst and T. van der Weide, A formal derivation of Heaps' Law, Information Sciences **170**, 263 (2005).
[29] I. Eliazar, The growth statistics of zipfian ensembles: Beyond heaps' law, Physica A: Statistical Mechanics and its Applications **390**, 3189 (2011).
[30] S. Bernhardsson, L. E. C. da Rocha, and P. Minnhagen, The meta book and size-dependent properties of written language, New Journal of Physics **11**, 123015 (2009).
[31] L. Lü, Z.-K. Zhang, and T. Zhou, Zipf's Law leads to Heaps' Law: Analyzing their relation in finite-size systems, PLOS ONE **5**, 1 (2010).
[32] T. M. Apostol, An Elementary View of Euler's Summation Formula, The American Mathematical Monthly **106**, 409 (1999).