SUPPLEMENTARY MATERIAL

A. PERCENTAGE DISTRIBUTION OF ARTICLES IN THE WOS SYSTEM BY THE NUMBER OF SUBJECT CATEGORIES TO WHICH THEY ARE ASSIGNED

Number of categories	Distinct articles	%	Articles in the extended count	%
1	1,982,353	54.8	1,982,353	33.3
2	1,084,414	30.0	2,168,828	36.5
3	429,765	11.9	1,289,295	21.7
4	94,135	2.6	376,540	6.3
5	15,151	0.4	75,755	1.3
6	8,627	0.2	51,762	0.9

B. A NUMERICAL EXAMPLE OF THE CITATION DISTRIBUTIONS DISCUSSED IN THE TEXT

1. The G8 classification system

Assume that there are eleven distinct articles classified in four clusters in the G8 system, say $\boldsymbol{c_j}$, j = 1, 2, 3, 4:

$\boldsymbol{c_1}$ = (0, 3, 12) with $\mu_1 = 15/3 = 5$,
$\boldsymbol{c_2}$ = (1, 2, 6) with $\mu_2 = 9/3 = 3$,
$\boldsymbol{c_3}$ = (4, 8) with $\mu_3 = 12/2 = 6$,
$\boldsymbol{c_4}$ = (5, 7, 9) with $\mu_4 = 21/3 = 7$.

The overall citation distribution under the G8 system, $\boldsymbol{C} = \cup_j \{\boldsymbol{c_j}\}$, is the following:

$\boldsymbol{C}$ = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 12).

The normalized cluster distributions are:

$\boldsymbol{c^*_1} = (0/\mu_1, 3/\mu_1, 12/\mu_1) = (0, 3/5, 12/5)$,
$\boldsymbol{c^*_2} = (1/\mu_2, 2/\mu_2, 6/\mu_2) = (1/3, 2/3, 2)$,
$\boldsymbol{c^*_3} = (4/\mu_3, 8/\mu_3) = (2/3, 4/3)$,
$\boldsymbol{c^*_4} = (5/\mu_4, 7/\mu_4, 9/\mu_4) = (5/7, 1, 9/7)$.

The overall G8-normalized citation distribution under the G8 system, $\boldsymbol{C^*} = \cup_j \{\boldsymbol{c^*_j}\}$, is the following:

$\boldsymbol{C^*}$ = (0, 1/3, 3/5, 2/3, 2/3, 5/7, 1, 4/3, 9/7, 2, 12/5)

with $|\boldsymbol{C^*}| = |\boldsymbol{C}| = 11$.

2. The G6 classification system

Assume that, in addition to the previous eleven distinct articles, there are two more articles with 0 and 11 citations in system G6. Assume also that the thirteen distinct articles are classified in two clusters, say $\boldsymbol{d_g}$, g = 1, 2:

$\boldsymbol{d_1}$ = (0, 1, 2, 7, 9, 11) with $M_1 = 30/6 = 5$,
$\boldsymbol{d_2}$ = (0, 3, 4, 5, 6, 8, 12) with $M_2 = 38/7$.

The overall citation distribution under the G6 system, $\boldsymbol{D} = \cup_g \{\boldsymbol{d_g}\}$, is the following:

$\boldsymbol{D}$ = (0, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12).

The normalized cluster distributions are:

$\boldsymbol{d^*_1}$ = (0, 1/5, 2/5, 7/5, 9/5, 11/5),
$\boldsymbol{d^*_2}$ = (0, 21/38, 14/19, 35/38, 21/19, 28/19, 42/19).

The overall G6-normalized citation distribution under the G6 system, $\boldsymbol{D^*} = \cup_g \{\boldsymbol{d^*_g}\}$, is the following:

$\boldsymbol{D^*}$ = (0, 0, 1/5, 2/5, 21/38, 14/19, 35/38, 21/19, 7/5, 28/19, 9/5, 11/5, 42/19)

with $|\boldsymbol{D^*}| = |\boldsymbol{D}| = 13$.

3. The construction of the citation distribution $\boldsymbol{C^{**}}$

In the construction of the citation distribution $\boldsymbol{C^{**}}$ to assess the G6-normalization procedure using the G8 classification system for evaluation purposes, the four cluster citation distributions $\boldsymbol{c^{**}_j}$ are as follows:

$\boldsymbol{c^{**}_1}$ = (0, 21/38, 42/19),
$\boldsymbol{c^{**}_2}$ = (1/5, 2/5, 21/19),
$\boldsymbol{c^{**}_3}$ = (14/19, 28/19),
$\boldsymbol{c^{**}_4}$ = (35/38, 7/5, 9/5).

The overall G6-normalized citation distribution under the G8 system, $\boldsymbol{C^{**}} = \cup_j \{\boldsymbol{c^{**}_j}\}$, is the following:

$\boldsymbol{C^{**}}$ = (0, 1/5, 2/5, 21/38, 14/19, 35/38, 21/19, 7/5, 28/19, 9/5, 42/19).

Since there are two distinct articles in citation distribution $\boldsymbol{D^*}$ that do not belong to $\boldsymbol{C^{**}}$, we have $|\boldsymbol{C^{**}}| = 11 < |\boldsymbol{D^*}| = 13$.

4. The construction of the citation distribution $\boldsymbol{D^{**}}$

To assess the performance of the G8-normalization procedure using the G6 system for evaluation purposes, consider again the intersection $\boldsymbol{C} \cap \boldsymbol{D}$. For every article v in this set with citations d_{gv} in cluster g in system G6, there exists a distinct article l in some cluster j with the same number of citations, i.e. $c_{jl} = d_{gv}$. The normalized score of this article according to the G6 system becomes $d^{**}_{gv} = c_{jl}/\mu_j$, where μ_j is the mean citation in cluster j. In this way, we construct an overall G8-normalized citation distribution under the G6 system $\boldsymbol{D^{**}} = \cup_g \{\boldsymbol{d^{**}_g}\}$. Since $c_{jl}/\mu_j = c^*_{jl}$, so that $d^{**}_{gv} = c^*_{jl}$, every article in $\boldsymbol{D^{**}}$ belongs to $\boldsymbol{C^*}$. If there are some articles in the G8 system that belong to a small cluster with less than 250 articles in the G6 system, then the citation distribution $\boldsymbol{D^{**}}$ is strictly contained in $\boldsymbol{C^*}$. In turn, since there are some articles in the G6 system that belong to a cluster with less than 250 articles in the G8 system, we have $\boldsymbol{D^{**}} \subset \boldsymbol{C^*} \subset \boldsymbol{D^*}$, so that $|\boldsymbol{D^{**}}| < |\boldsymbol{C^*}| = 3.4$ million articles $< |\boldsymbol{D^*}| = 3.6$ million articles.

In the example, the two cluster citation distributions $\boldsymbol{d^{**}_g}$ are as follows:

$\boldsymbol{d^{**}_1}$ = (0, 1/3, 2/3, 1, 9/7),
$\boldsymbol{d^{**}_2}$ = (3/5, 2/3, 5/7, 4/3, 2, 12/5).

Since for each article in $\boldsymbol{D}$ there exists a distinct article in $\boldsymbol{C}$, the overall G8-normalized citation distribution under the G6 system, $\boldsymbol{D^{**}} = \cup_g \{\boldsymbol{d^{**}_g}\}$, coincides with $\boldsymbol{C^*}$, so that

$\boldsymbol{D^{**}}$ = (0, 1/3, 3/5, 2/3, 2/3, 5/7, 1, 4/3, 9/7, 2, 12/5)

with $|\boldsymbol{D^{**}}| = |\boldsymbol{C^{*}}| = 11 < |\boldsymbol{D^{*}}| = 13$.

5. The extended count in the example

Assume that there are three WoS categories, and that three distinct articles in $\boldsymbol{C}$ and $\boldsymbol{D}$ are assigned to several categories: the two articles with five and seven citations are assigned to the first and the third categories, while the article with three citations is assigned to the three categories. We assume that the seventeen articles in the extended count are distributed as follows:

$\boldsymbol{e_1} = (1, 3, 5, 7, 11)$ with $\mu_1 = 27/5$,
$\boldsymbol{e_2} = (0, 0, 3, 4, 6, 8)$ with $\mu_2 = 21/6$,
$\boldsymbol{e_3} = (2, 3, 5, 7, 9, 12)$ with $\mu_3 = 38/6 = 19/3$.

The overall extended citation distribution under the WoS system, $\boldsymbol{E} = \cup_k \{\boldsymbol{e_k}\}$, is the following:

$\boldsymbol{E} = (0, 0, 1, 2, 3, 3, 3, 4, 5, 5, 6, 7, 7, 8, 9, 11, 12)$

The normalized extended category citation distributions are:

$\boldsymbol{e^{*}_1} = (5/27, 15/27, 25/27, 35/27, 55/27)$,
$\boldsymbol{e^{*}_2} = (0, 0, 18/21, 24/21, 36/21, 48/21)$,
$\boldsymbol{e^{*}_3} = (6/19, 9/19, 15/19, 21/19, 27/19, 36/19)$.

The overall WoS-normalized citation distribution under the WoS system, $\boldsymbol{E^{*}} = \cup_k \{\boldsymbol{e^{*}_k}\}$, is the following:

$\boldsymbol{E^{*}} = (0, 0, 9/19, 5/27, 15/27, 6/19, 18/21, 25/27, 21/19, 24/21, 35/27, 27/19, 15/9, 36/21, 36/19, 55/27, 48/21)$

with $|\boldsymbol{E^{*}}| = |\boldsymbol{E}| = 17$.

6. The construction of the $\boldsymbol{E^{**}}$ citation distribution

In order to assess the performance of the G8-normalization procedure when the WoS system is used for evaluation purposes, consider the intersection between $\boldsymbol{C}$ and the set of distinct articles in $\boldsymbol{E}$. Consider an article u in this set with citations c_{ju} in cluster j in system G8. Assume, for example, that there are two articles v and w in categories l and h in the WoS system with this same number of citations, i.e. $e_{lv} = e_{hw} = c_{ju}$. In this case, the distinct article with raw citations c_{ju} will give rise to two normalized articles according to the WoS system, say e^{**}_{lv} and e^{**}_{hw}. Since the mean citation in cluster j is μ_j, the two normalized articles become $e^{**}_{lv} = e^{**}_{hw} = c_{ju}/\mu_j$. These articles will be included in the extended normalized citation distributions $\boldsymbol{e^{**}_l}$ and $\boldsymbol{e^{**}_h}$. In this way, we construct a G8-normalized overall citation distribution under the WoS system $\boldsymbol{E^{**}} = \cup_k \{\boldsymbol{e^{**}_k}\}$. Note, however, that $c_{ju}/\mu_j = c^{*}_{ju}$, so that $e^{**}_{jv} = e^{*}_{kv}$, and $c^{**}_{jw} = e^{*}_{hw}$. Therefore, the citation distribution $\boldsymbol{C^{*}}$ is strictly contained in $\boldsymbol{E^{**}}$. On the other hand, since there are some articles in the WoS system that belong to a cluster with less than 250 articles in the G8 system, $\boldsymbol{E^{**}}$ is strictly contained in $\boldsymbol{E^{*}}$. Therefore, we have $|\boldsymbol{C^{*}}|$ = 3.4 million articles < $|\boldsymbol{E^{**}}| < |\boldsymbol{E^{*}}|$ = 5.9 million articles.

In the example, the three category citation distributions $\boldsymbol{e^{**}_k}$ are as follows:

$\boldsymbol{e^{**}_1} = (1/3, 3/5, 5/7, 1)$,
$\boldsymbol{e^{**}_2} = (0, 3/5, 2/3, 4/3, 2)$,
$\boldsymbol{e^{**}_3} = (3/5, 2/3, 5/7, 1, 9/7, 12/5)$.

The extended overall G8-normalized citation distribution under the WoS system, $\boldsymbol{E^{**}} \cup_k \{\boldsymbol{e^{**}_k}\}$, becomes

$\boldsymbol{E^{**}} = (0, 1/3, 3/5, 3/5, 3/5, 2/3, 2/3, 5/7, 5/7, 1, 1, 4/3, 9/7, 2, 12/5)$.

Since there are two distinct articles in citation distribution $\boldsymbol{E^*}$ that do not belong to $\boldsymbol{E^{**}}$, we have $|\boldsymbol{E^{**}}| = 15 < |\boldsymbol{E^*}| = 17$.

7. The construction of the $\boldsymbol{E^{***}}$ citation distribution

In order to assess the performance of the G6-normalization procedure when the WoS system is used for evaluation purposes, we proceed in a similar way. In other words, we construct a G6-normalized overall citation distribution under the WoS system $\boldsymbol{E^{***}} = \cup_k \{\boldsymbol{e^{***}_k}\}$. In this case, we have $|\boldsymbol{D^*}|$ = 3.6 million articles < $|\boldsymbol{E^{***}}| < |\boldsymbol{E^*}|$ = 5.9 million articles.

In the example, the three category citation distributions $\boldsymbol{e^{***}_k}$ are as follows:

$\boldsymbol{e^{***}_1}$ = (1/5, 21/38, 35/38, 7/5, 11/5),
$\boldsymbol{e^{***}_2}$ = (0, 0, 21/38, 28/38, 21/19, 28/19),
$\boldsymbol{e^{***}_3}$ = (2/19, 21/38, 35/38, 7/5, 9/5, 42/19).

The extended overall G6-normalized citation distribution under the WoS system, $\boldsymbol{E^{***}} \cup_k \{\boldsymbol{e^{***}_k}\}$, becomes

$\boldsymbol{E^{***}}$ = (0, 0, 1/5, 2/5, 21/38, 21/38, 21/38, 14/19, 35/38, 35/38, 21/19, 7/5, 7/5, 28/19, 9/5, 11/5, 42/19).

Since for any distinct article in citation distribution $\boldsymbol{E^*}$ there exists a distinct article in $\boldsymbol{D^*}$, we have $|\boldsymbol{E^{***}}| = |\boldsymbol{D^*}| = |\boldsymbol{E^*}| = 17$.

8. The construction of the $\boldsymbol{C^{***}}$ citation distribution

In order to assess the performance of the WoS-normalization procedure when the G8 system is used for evaluation purposes, consider the intersection between $\boldsymbol{C}$ and the set of distinct articles in $\boldsymbol{E}$. Note that for every distinct article u in this set receiving c_{ju} citations in the cluster citation distribution $\boldsymbol{c_j}$, there must be one or more articles in the original WoS system with the same number of citations. Assume, for example, that there are two articles v and w in categories k and h, respectively, with $e_{kv} = e_{hw} = c_{ju}$. Since the mean citations in categories k and h are μ_k and μ_h, the two normalized articles according to the G8 system become $c^{***}_{jv} = c_{ju}/\mu_k$ and $c^{***}_{jw} = c_{ju}/\mu_h$. Both of these articles will be included in the extended normalized citation distribution $\boldsymbol{c^{***}_j}$. In this way, we construct a WoS-normalized overall citation distribution under the G8 system $\boldsymbol{C^{***}} = \cup_j \{\boldsymbol{c^{***}_j}\}$. Note, however, that $c_{ju}/\mu_k = e_{kv}/\mu_k = e^*_{kv}$ and $c_{ju}/\mu_k = e_{hw}/\mu_h = e^*_{hw}$, so that $c^{***}_{jv} = e^*_{kv}$, and $c^{***}_{jw} = e^*_{hw}$. Therefore, every element in $\boldsymbol{C^{***}}$ belongs to the overall extended citation distribution $\boldsymbol{E^*}$. Since there are some articles in the WoS system whose corresponding distinct articles belong to a small cluster with less than 250 articles, the citation distribution $\boldsymbol{C^{***}}$ is strictly contained in $\boldsymbol{E^*}$, that is, $\boldsymbol{C^{***}} \subset \boldsymbol{E^*}$. On the other hand, since for any distinct article in $\boldsymbol{C}$ there might be two or more articles in the WoS the citation distribution $\boldsymbol{C^{***}}$ strictly contains $\boldsymbol{C^*}$, so that $|\boldsymbol{C}|$ = 3.4 million articles < $|\boldsymbol{C^{***}}| < |\boldsymbol{E^*}|$ = 5.9 million articles.

In the example, the four cluster citation distributions $\boldsymbol{c^{***}_j}$ are as follows:

$\boldsymbol{c^{***}_1}$ = (0, 9/19, 15/27, 18/21, 36/19),
$\boldsymbol{c^{***}_2}$ = (5/27, 6/19, 36/21),
$\boldsymbol{c^{***}_3}$ = (24/21, 16/7),
$\boldsymbol{c^{***}_4}$ = (25/27, 15/19, 21/19, 35/27, 27/19).

The overall WoS-normalized citation distribution under the G8 system, $\boldsymbol{C^{***}} = \cup_j \{\boldsymbol{c^{***}_j}\}$, is the following:

$\boldsymbol{C^{***}}$ = (0, 9/19, 5/27, 15/27, 6/19, 15/19, 18/21, 25/27, 21/19, 24/21, 35/27, 27/19, 12/7, 36/19, 16/7).

Since there are two distinct articles in citation distribution $\boldsymbol{E^*}$ that do not belong to $\boldsymbol{C^{***}}$, we have $|\boldsymbol{C^{***}}| = 15 < |\boldsymbol{E^*}| = 17$.

9. The construction of the $\boldsymbol{D^{***}}$ citation distribution

In order to assess the performance of the WoS-normalization procedure when the G6 system is used for evaluation purposes, consider the intersection between $\boldsymbol{D}$ and the set of distinct articles in $\boldsymbol{E}$. For every distinct article u in this set receiving d_{gu} citations in the cluster citation distribution $\boldsymbol{d_g}$ in the G6 system, there must be one or more articles in the WoS system with the same number of citations. Assume, for example, that there are two articles v and w in categories k and h with $e_{kv} = e_{hw} = d_{gu}$. In this case, the distinct article with raw citations d_{gu} will give rise to two normalized articles according to the WoS system, say d^{***}_{gv} and d^{***}_{gw}. Since the mean citations in categories k and h are μ_k and μ_h, the two normalized articles become $d^{***}_{gv} = e_{kv}/\mu_k = e^*_{kv}$ and $d^{***}_{gw} = e_{hw}/\mu_h = e^*_{hw}$. Both of these articles will be included in the extended normalized citation distribution $\boldsymbol{d^{***}_g}$. In this way, we construct a WoS-normalized overall citation distribution under the G6 system $\boldsymbol{D^{***}} = \cup_g \{\boldsymbol{d^{***}_g}\}$. The citation distribution $\boldsymbol{D^{***}}$ is strictly contained in $\boldsymbol{E^*}$, and strictly contains $\boldsymbol{D^*}$, so that $|\boldsymbol{D}| = 3.6$ million articles $< |\boldsymbol{D^{***}}| < |\boldsymbol{E^*}| = 5.9$ million articles.

In the example, the two cluster citation distributions $\boldsymbol{d^{***}_g}$ are as follows:

$\boldsymbol{d^{***}_1}$ = (0, 5/27, 6/19, 35/27, 21/19, 27/19, 55/27),
$\boldsymbol{d^{***}_2}$ = (0, 15/27, 18/21, 9/19, 24/21, 25/27, 15/19, 36/21, 48/21, 36/19).

Since for any article in citation distribution $\boldsymbol{E}$ there exists a distinct article in citation distribution $\boldsymbol{D}$, the overall WoS-normalized citation distribution under the G6 system, $\boldsymbol{D^{***}} = \cup_j \{\boldsymbol{d^{***}_j}\}$, coincides with $\boldsymbol{E^*}$, so that

$\boldsymbol{D^{***}}$ = (0, 0, 9/19, 5/27, 15/27, 6/19, 18/21, 25/27, 21/19, 24/21, 35/27, 27/19, 15/9, 36/21, 36/19, 55/27, 48/21)

with $|\boldsymbol{D^{***}}| = |\boldsymbol{E^*}| = 17$.

C. A METHOD TO EVALUATE THE DIFFERENCES BETWEEN A PAIR OF CLASSIFICATION SYSTEMS

Let $\boldsymbol{x_j}$, $\boldsymbol{x_g}$, and $\boldsymbol{x_k}$ be the sets of the top X% most cited articles in cluster citation distributions $\boldsymbol{c_j}$ and $\boldsymbol{c_g}$, and category citation distribution $\boldsymbol{e_k}$. Denote the union of these sets by $\boldsymbol{X^{G8}} = \cup_j \{\boldsymbol{x_j}\}$, $\boldsymbol{X^{G6}} = \cup_g \{\boldsymbol{x_g}\}$, and $\boldsymbol{X^{WoS}} = \cup_k \{\boldsymbol{x_k}\}$. There are two ways of comparing the WoS system with systems G8 or G6: in terms of distinct articles, or in terms of extended articles. Let us begin by defining the set $\boldsymbol{X^{WoS}}$(Dist) of distinct articles in $\boldsymbol{X^{WoS}}$. We can now form the following two intersections of distinct articles:

$$\boldsymbol{X^{W8}}(\text{Dist}) = \boldsymbol{X^{WoS}}(\text{Dist}) \cap \boldsymbol{X^{G8}}$$

and

$$\boldsymbol{X^{W6}}(\text{Dist}) = \boldsymbol{X^{WoS}}(\text{Dist}) \cap \boldsymbol{X^{G6}}.$$

The difference between the top X% most cited articles in the G8 and WoS systems can be measured by the percentage that the articles in $\boldsymbol{X^{G8}} - \boldsymbol{X^{W8}}$(Dist) represent in $\boldsymbol{X^{G8}}$. Similarly, the difference between the top X% most cited articles in the G6 and WoS systems can be measured by the percentage that the articles in $\boldsymbol{X^{G6}} - \boldsymbol{X^{W6}}$(Dist) represent in $\boldsymbol{X^{G6}}$. The results for these two expressions are presented in Tables 2.a and 2.b under the heading "Difference in distinct articles".

On the other hand, let $\boldsymbol{X^{G6}}$(Ext) and $\boldsymbol{X^{G8}}$(Ext) be the sets of extended articles in $\boldsymbol{X^{G6}}$ and $\boldsymbol{X^{G8}}$. We can now form the following two intersections of extended articles:

$$X^{W8}(\text{Ext}) = X^{WoS} \cap X^{G8}(\text{Ext})$$

and

$$X^{W6}(\text{Ext}) = X^{WoS} \cap X^{G6}(\text{Ext}).$$

The difference between the top X% most cited articles in the G8 and WoS systems can be measured by the percentage that the articles in $X^{WoS} - X^{W8}(\text{Ext})$ represents in X^{WoS}. Similarly, the difference between the top X% most cited articles in the G6 and WoS systems can be measured by the percentage that the articles in $X^{WoS} - X^{W6}(\text{Ext})$ represents in X^{WoS}. The results for these two expressions are presented in Tables 2.a and 2.b under the heading "Difference in extended articles".

For the comparison between the G6 and G8 systems, let X^{68} be the set of distinct articles common to both systems, namely, let $X^{68} = X^{G6} \cap X^{G8}$. There are two ways of measuring the difference between the top X% most cited articles in both systems: through the percentage that the articles in $X^{G8} - X^{68}$ represent in X^{G8}, and through the percentage that the articles in $X^{G6} - X^{68}$ represent in X^{G6}. The results for these two expressions are presented under the headings "Difference in terms of the G8 system" and "Difference in terms of the G6 system" in Table 2.c.