Credit rating has spawned a huge industry, far more complex than the simple counting game at its core. Still, uncertainty pervades ratings and deserves more disclosure. This needn’t make ratings gyrate wildly. In analogy to magnetism, a mild preference for agreement on relative safety can induce remarkable stability. The flip side is that a few defaults can trigger widespread ratings shifts.

The marvel is less that public confidence in Western finance occasionally plummets than that so much persists and rebuilds. It’s like a starfish growing back arms. One of the confidence boosters is credit rating.

Credit rating proper is associated with regulators and with major credit rating agencies like Moody’s and Standard & Poor’s (S&P). By convention, credits are assigned letter grades. Triple-A (Aaa for Moody’s, AAA for S&P) is the highest rating, followed by double-A, single-A, triple-B (or Baa), and so on down to C. Grades below triple-A are subdivided by number or plus-or-minus sign and sometimes attach warnings of imminent migration.

In general, any ranking of credit safety, including the ranking implicit in trading, can be considered a credit rating. Nearly every major financial institution has credit scoring departments or uses third-party scoring services. Some services, like Riskmetrics or CreditRisk+, provide a mixture of scores and advice on scoring methodology.

This chapter will examine the ratings process in more depth. We will find that the counting game isn’t so simple after all, and warrants respect. However, it pretends to more precision than it can deliver. Much of the agreement we observe about risk is just a form of social bonding.
While useful in encouraging joint effort, it occasionally bodes huge misdirection.

Credit Grades

Credit rating is a valuable service. It takes time and expertise to identify relevant signals and estimate their implications. Having invested time and gained expertise, it makes sense to resell the estimates to others. Debtors appreciate knowing where they stand. Lenders appreciate the feedback on their own evaluations.

To some extent ratings are self-fulfilling, as high ratings help issuers roll over debt while low ratings hinder. Some institutions are restricted by charter to holding investment-grade debt, rated triple-B or higher. Sub-investment grades were once known mostly as “junk” and could hardly place bonds at all. Nowadays markets actively trade a broad spectrum of credit. Their feedback provides a check on ratings, with wide discrepancies encouraging reassessment.

Credit grades don’t claim to indicate the absolute default risk. The latter fluctuates with economic cycles, technological change, and political shifts outside the scope of most rating exercises. Also, the rating agencies don’t want to be held culpable for potential debt implosions. They emphasize that their grades are opinions on relative rankings within asset classes, with potentially limited relevance across asset classes or over time. Moody’s (Moody’s Investor Service 2009) reference on rating symbols and definitions doesn’t mention a single quantitative benchmark.

Fortunately for the rating agencies, regulators don’t believe their disavowals. They use the credit grades, or let the entities they regulate use the credit grades, as proxies for absolute default risk. To facilitate comparison, major rating agencies periodically report the empirical default rates for different credit grades over horizons ranging from one year to more than ten years.

Table 8.1 is a compilation from Moody’s (Emery et al. 2009: exhibit 36) of average one-year U.S. corporate default rates between 1920 and 2008. They are grouped by Moody’s credit grade at the beginning of the year in which default occurred. Next to it is similar data from S&P (Vazza, Aurora, and Kraemer 2010: table 25) covering 1980 through 2009.

The two lists are remarkably similar given the differences in agency and time period. The only major discrepancy occurs in the lowest categories,
where Moody’s shows much fewer defaults. That discrepancy would largely vanish if Moody’s averages were calculated from 1980 onward. Compilations from different asset classes, countries, and periods show more diversity. However, several regularities stand out:

- Default rates are generally low except in the lowest credit rungs or sharp recessions. To qualify as investment grade, default risk over the economic cycle should be less than 50 bps per annum, even before we factor in typical recovery rates of roughly half.
- Credit risks scale logarithmically, with a drop in letter grade roughly quadrupling the one-year default risk. Since estimation errors and intrayear regime switches muddy the actual results, the target gradient must be even steeper. Writing for Moody’s, Yoshizawa (2003) describes the “idealized” default risks as roughly doubling per drop in alphanumeric notch and hence octupling per drop in primary letter grade.
- Macroeconomic fluctuations can swing the default risk a letter grade’s worth in either direction. Hence, credit grades point to average default rates over a cycle.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaa</td>
<td>0</td>
<td>AAA</td>
<td>0</td>
</tr>
<tr>
<td>Aa</td>
<td>6</td>
<td>AA</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>9</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>Baa</td>
<td>27</td>
<td>BBB</td>
<td>28</td>
</tr>
<tr>
<td>Ba</td>
<td>106</td>
<td>BB</td>
<td>104</td>
</tr>
<tr>
<td>B</td>
<td>340</td>
<td>B</td>
<td>486</td>
</tr>
<tr>
<td>Caa-C</td>
<td>1,325</td>
<td>CCC-C</td>
<td>2,892</td>
</tr>
</tbody>
</table>
Factor Analysis

For good credit analysis, it is crucial to identify a few common driving factors and to collect a lot of relatively independent observations on their impact. This makes U.S. corporate debt an inviting subject of study. There are thousands of different issuers, with tens of thousands of years of debt servicing experience under a relatively stable legal environment.

Using Fisher’s (1936) discriminant analysis, Altman (1968) developed a famous credit score based on working capital, retained earnings, earnings before interest and taxes, and sales—all measured relative to assets—and net worth divided by liabilities. More refined models have been developed since. Servigny and Renault (2004) provide an excellent overview.

To keep scores from predicting a negative default risk, models typically estimate the log odds. In the range of most interest, log odds basically coincide with log risk. This helps justify the association of credit grades with a logarithmic scale of default risk.

Rating agencies combine quantitative scoring with discretionary judgments. They also take into account market information like prices, volatility, and credit spreads. Their primary focus, however, is economic drivers, since they aim to provide independent guidance on fair price.

From a counting game perspective, viewing different credits as bundles of a smaller set of primary factors raises the effective $D$ and $T$. This makes estimates more reliable and trims uncertainty. Without these kinds of decompositions, formal or informal, credit rating would be a farce.

Even with a factor analysis approach, we don’t get nearly as many plausibly independent-but-relevant observations as we would like. Look at the heavy clustering in Figure 8.1. Inspired by Giesecke et al. (2010), it depicts yearly default rates from 1866 through 2009 for bonds issued by U.S. non-financial corporations.

The default rates are strikingly uneven. They fluctuate in irregular cycles of varying intensity, with peaks about ten times neighboring troughs. Intensity has receded overall, though not in linear fashion. One can plausibly argue that regime changes occurred around 1900, 1940, and 1980.

The credit cycles do not coincide with standard macroeconomic cycles. Giesecke and colleagues find a correlation of 0.25. If we use the fluctuations themselves to define an explanatory factor, we are bound to overfit and have limited predictive power. If we ignore the fluctuations, our observations won’t be conditionally independent. If we search for cycle proxies, we’ll experience both problems, though it is hoped in smaller measure.
Hence, even for the bond market as a whole it’s hard to identify a long-term default rate within a factor of two. The similarities in Table 8.1 between 1920–2006 averages and 1980–2009 averages seem a well-chosen coincidence. Judging from Figure 8.1, any other long period would show different averages.

Even if we correctly identify the mean default rate, estimated default rates at extremes can be unreliable. If we fit a straight line through a cluster of data, errors in the slope affect the outer estimates much more than the inner. Nor can we be sure that the best fit is a straight line, and again the errors matter most in the extremes. It’s always riskier to extrapolate than to interpolate.

This risk comes to the fore whenever safe assets meet severe crisis. We’ll likely be uncertain about both safety and severity; the outcome will refine our views. The main reason U.S. mortgages retained such high ratings until the national housing bubble burst is that they had never previously experienced a bubble burst of that scale.
Grading Uncertainty

Credit ratings summarize decades of detailed analysis and the web of confidence they support. That doesn’t mean any particular credit risk is graded correctly. The raters can’t easily know, and we can’t easily check.

For example, the post-1920 historical averages suggest that A-grade default risk should be less than 16 bps. That’s less than one default per 600 years of servicing. But a 20 bps risk could plausibly stay default-free that long. Indeed, a 20 bps risk has a 1% chance of servicing for 2,300 years straight.

Higher credit grades are even harder to distinguish. Suppose we think a credit has either no risk or 2 bps of risk and accordingly warrants a triple-A or double-A grade. Starting from even odds, we would need the equivalent of 20,000 years of unbroken servicing to become 99% confident in the triple-A label.

It’s safe to say that no one has ever conclusively identified triple-A safety. Any risk that low would be overwhelmed by invasions, volcano eruptions, plagues, civil wars, fraud, terrorism, mass extinctions, or other disasters we can’t yet imagine. I am tempted to call it “Who Knows What Can Happen” risk. A refined name is “force majeure.”

The only plausible way to defend particular credit labels is to identify them as hopeful means of much broader confidence intervals. To assist that, rating agencies ought to report the confidence interval they think applies. Analytically this corresponds to reporting $T$ as well as $E = \frac{D}{T}$. But expressing sample limitations and other uncertainties in terms of upper and lower grades will make them easier to digest.

Reporting confidence intervals will make it easier to compare risk across asset classes. Sovereign credit risk is a lot harder to identify than corporate credit risk, as there are many fewer observations and the mechanisms for enforcing claims are murkier. Broader confidence intervals can highlight this. Confidence intervals should also be wide for poorly documented securitizations or credits facing unprecedented stresses.

Confidence intervals will also assist in pricing longer-dated bonds. The more dispersed beliefs are around the mean, the more a long-dated payment tends to be worth. This may surprise those who expect risk-neutral investors not to care about uncertainty and risk-averse investors to loathe it. But technically it is analogous to the value of interest rate convexity, a concept well known in bond pricing. For an extreme example, compare these two scenarios:
- 99% chance of being riskless and a 1% chance of being worthless
- 100% chance of 1% annual default risk

A bond payment ten years forward needs be discounted only 1% for credit risk in the first case, versus nearly 10% in the second.

Change won’t come easy. The current system is well ensconced. Rating agencies have been awarding single grades for generations. Regulators have grown accustomed to treating risk estimates as certain. Neither side has much expertise in analyzing uncertainty systematically. Neither side seems itching to gain it.

Nevertheless, high/low grading offers some advantages to both rating agencies and regulators. By allowing for mixed ratings, it can soften the uproar caused by sudden loss of investment-grade status. It encourages more timely adjustment of ratings, since agencies can visibly qualify their decisions. It can tidy up the credit watches and advisories that hint at uncertainty.

On the regulatory side, formal disclosures of uncertainty can aid in setting contingent reserves. While Basel II wanted banks to develop their own internal ratings system, the requirements were cumbersome. Many banks found it cheaper to purchase ratings from rating agencies, as Basel II allowed. In effect Basel II delegated a lot of regulatory authority to rating agencies, without demanding much from rating agencies in return. More disclosure of uncertainty is a wholly legitimate demand.

Some critics would go farther and remove rating agencies’ special status as regulatory advisors. I disagree. Overall, the delegation breeds more competence and less corruption than wholly government-run ratings. We do, however, need to trim some ludicrous incentives to shade risk estimates down. Demanding more disclosure will help, though not as much as trimming the marginal leverage for higher credit grades.

A prominent example is the triple-A rating accorded U.S. and U.K. sovereign debt despite agency warnings about mounting servicing burdens. By rating agencies’ own criteria, even a strong hint of problems should nix a triple-A grade. But of course sovereigns are special, both in their funding possibilities and in their ability to take away quasi-regulatory privileges. Even the recent downgrades of PIIGS, long overdue, triggered a huge outcry. High/low grading would provide a quieter middle ground.

Uncertainty grading is no panacea. The housing bubble had a constellation of other causes, ably surveyed in Ellis (2008). The overstretch of sovereign debt in the United States and Europe mostly stems from excessive...
benefits to pensioners and public employees. Deflating overconfidence in servicing won’t make problems go away. But it will expose them better to public view. Finance needs to get away from the mindset that others’ irresponsibility excuses one’s own.

Credit Migration

Credit ratings are bound to migrate because of estimation noise or actual regime switches. The previously mentioned Moody’s and S&P studies paint similar pictures of ratings migration:

- Averaging over the past quarter century, roughly 25% of corporate alphanumeric ratings change per year. Some 40% of these, or 10% of the total, change primary letter grades.
- Migrations are roughly 50% more frequent in subinvestment grades than in investment grades.
- Most migrations head to neighboring ratings. Occasionally there are big jumps, especially downgrades related to defaults.
- Downgrades are more frequent than upgrades for investment-grade credits. Credits below B tend to migrate up if they don’t default. Credits between single-B and double-B migrate relatively equally in each direction.
- Migration rates are uneven. Downgrades surge around actual default surges. Upgrades surge when defaults recede. However, the fluctuations are not nearly as sharp for migration rates as for actual defaults.

Bayesian updating for evidence of default helps explain the jumps and irregularities. To model the gradual diffusion and mean reversion, let’s step back a moment and think about default risks the way physicists think about gases or biologists think about populations. Some factors make them expand, some factors make them contract, and some just add noise. Expansion outweighs when default risks are low, because money is easy and borrowers get complacent. Contraction outweighs when default risks are high, because borrowers work down their burdens through fiscal tightening or restructuring.

To simplify calculations, imagine these forces operating continuously over time without jumps. Since default risk can’t be negative, let’s model
noise as proportional to a positive power $m$ of current risk $\theta$. A tractable model that delivers this is

$$d\theta = a\theta^{2m-1}(1 - b\theta)dt + c\theta^m dz,$$  

(8.1)

where $a$, $b$, and $\sigma$ are constants, $d\theta$ is the change in default risk over infinitesimal time $dt$, and $dz$ represents standard Brownian motion with zero drift and unit rate of volatility. May (1973) used an equation like this with $m = 1$ to model a large population $\theta$ subject to a natural growth rate $a$, a stable carrying capacity $\frac{1}{b}$, and random influences $c$. He showed that it implied a long-term gamma distribution for $\theta$.

Indeed, equation (8.1) implies $\theta$ is long-term gamma distributed for every $m$, unless there’s no interior equilibrium at all. Examples include the famed Cox, Ingersoll, and Ross (1985) model of the spot interest rate, where $m = \frac{1}{2}$. Moreover, Dennis and Patil (1984) have shown that most growth and decay models with nonnegativity constraints imply a long-term gamma distribution or something close, even when (8.1) does not exactly apply.

Hence, gamma distributions arise naturally both in the long-term distribution of default risk and the short-term distribution of beliefs about default risk. However, let me emphasize that the two gamma distributions aren’t nearly the same. The long-term shape parameter (i.e., the counterpart to $D$ in the model of credit beliefs) must significantly exceed one to account for the concentration of ratings around double-$B$.

The Appendix analyzes equation (8.1) in more detail. The most curious finding concerns the power $m$. Faster migration among lower ratings appears to require $m > 1$. Yet most credit models assume $m \leq 1$.

Estimation error makes the discrepancy even more striking. As we have seen, high ratings are much harder to estimate than low ratings. All else being equal, we should expect to see more migration among higher grades than among lower grades.

In short, top grades seem unusually sticky. Why? Presumably rating agencies are reluctant to publicly change their minds. However, that can’t be all there is to it, since the frequency of ratings changes varies significantly over time, between credit grades, and across asset classes. Also, the market has plenty of participants who watch the credit watchers. Systematically blatant errors should incite systematic countertrading, which I would expect to leave noticeable tracks.
Deference to Consensus

The rest of this chapter will explain ratings stickiness in terms of a preference for consensus. This may sound like either a tautology or a slur on humanity. However, I’m not imputing a strong preference, just a mild deference to “the wisdom of the crowd” when highly uncertain. And I will apply it in an unusual way, namely to show why top ratings are stickier than others.

The model is inspired by Ising (1925). Ising-type models are widely used in physics to explain phase transitions in matter. While highly stylized, their core predictions often match the predictions of far more complex models.

Consider, for example, ferromagnetism. Place a cool block of iron in even a weak magnetic field and the atoms will tend to align in spin, creating a magnet. Shut off the field and the magnetism persists. Heat the iron enough and the magnetism disappears. How can we reconcile these behaviors?

The basic Ising explanation is that the iron atoms face two kinds of forces. One is ordinary random motion related to heat, which favors maximum disorganization. The other is the attraction of like spins, which encourages neighbors to align in the same direction. At low temperatures only a few atoms will have enough energy to flip spins against the alignment, and these will randomly flip back, so the core alignment perpetuates itself.

Similar phenomena occur when water cools to ice. A crystal formed randomly encourages neighbors to coalesce with it. Ising-type models help explain how the transitions can be sharp even though the temperature change is gradual.

In our model, many observers make many comparisons between two credits. Each comparison estimates a difference $\omega$ in perceived default risk, which might be any real number. However, all that gets reported is which credit is riskier. This is a stylized version of ordinal credit rankings, with very fine gradations.

Why don’t the full $\omega$ values get reported? There are several plausible explanations, including:

- Spelling out the environmental assumptions underlying point estimates might require more attention than the estimates are worth.
- Since different $\omega$ likely assume different environments, comparing them would be difficult without conversion to coarser measures.
- Reporters might be embarrassed to report $\omega$ values way out of line with consensus.
We can view these reports as signals $S = \pm 1$, with rare exceptions where $S = 0$. A fully independent reporter should set $S$ to the sign of $\omega$. However, reporting usually involves some deference to consensus. This leads to a decision rule along the lines of

$$S(\omega) = \text{sgn}(\omega + \epsilon \cdot \text{consensus}),$$

(8.2)

where sgn indicates the sign and $\epsilon$ is a tiny positive weight.

Clearly, there must be a threshold $\bar{\omega} \equiv -\epsilon \cdot \text{consensus}$ where $S$ vanishes. Below that threshold $S$ is negative; above it $S$ is positive. Denoting the cumulative distribution of $\omega$ by $H$, and assuming $\bar{\omega}$ is rare, consensus should stabilize near $(1 - H(\bar{\omega})) \cdot 1 + H(\bar{\omega}) \cdot (-1) = 1 - 2H(\bar{\omega})$. In equilibrium,

$$H(\bar{\omega}) = \frac{1}{2} \left(1 + \frac{\bar{\omega}}{\epsilon}\right).$$

(8.3)

By Ising model standards, this is an extraordinarily simple equilibrium condition. But it generates similar kinds of results.

Ratings Stickiness

To demonstrate Ising-type behavior with minimal clutter, I will assume that the two credits have identically distributed beliefs. In that case $H(0) = \frac{1}{2}$, so one equilibrium sets $\bar{\omega} = 0$ as expected. I will also assume that the mode or peak density $h_{\text{max}}$ of $H$ occurs at zero, which should be true unless beliefs are distributed very unnaturally.

To check for other equilibria, Figure 8.2 graphs both sides of equation (8.3) as functions of candidate $\bar{\omega}$ values. The right-hand side, which I call the control line, has a steep slope of $\frac{1}{2\epsilon}$; the chart is stretched for clarity. If that slope exceeds $h_{\text{max}}$, the origin will be the only equilibrium.

Otherwise, there will be two additional equilibria. Where $H$ is left of the control line, the candidate $\bar{\omega}$ will tend to increase—i.e., more reports will flip negative. Where $H$ is to the right of the control line, the candidate $\omega_0$ will tend to decrease. This makes the outer equilibria stable and the equilibrium at zero unstable.

Hence, when $h_{\text{max}}$ is sufficiently high, one credit slides into favor and stays there absent a major upheaval. The biggest and stickiest distortions occur when most beliefs about $\omega$ are tightly concentrated.
around the origin. In that case $H$ resembles a step function (nearly flat, nearly vertical and nearly flat again), and the stable equilibria are close to $\pm \varepsilon$.

This is most likely to occur with top-rated credits. As seen in Chapter 6, their beliefs tend to cluster tightly near zero, with infinite density at zero itself. If we compare two such beliefs chosen at random, or two mean beliefs calculated with random exclusions, the resulting $h_{\text{max}}$ will be infinite as well. This will guarantee distortion.

While $\varepsilon$ is tiny, it can be large relative to top-rated default risks. We thus have a clear explanation for the stickiness of top ratings. A taste for consensus creates a self-sustaining bias. The paucity of news for such credits—i.e., they tend to service just as expected—helps keep the bias in place. Most observers perceive both risks as so tiny that they’re reluctant to challenge the accepted ranking.

The bias isn’t permanent. When evidence of default does arrive for top-rated credits, even indirect and fractional evidence, it tends to come as a huge shock. This shakes up rankings the way that a furnace shakes up ferromagnetism. When the risk environment cools, new rankings will emerge. Some of them will incorporate genuinely new information. Some may reflect just another random consensus.

Figure 8.2
Equilibria for $\bar{\omega}$ with Symmetric Beliefs

Downloaded from cupola.columbia.edu
Our analysis also applies to stickiness of perceptions about fiat money, wealth, and debt sustainability. Between deference to consensus and the infrequency of payment crises, people can easily convince each other that the real risks are negligible. The network of trust can shave the real risks by facilitating rollover and encouraging investment. Nevertheless, the model here reminds us that our strongest beliefs can have weak foundations.

“What a disgrace this worship of credit ratings is,” said Pandora. “Nothing good can come of idolatry.”

“Pygmalion would beg to differ,” said Prometheus. “The statue he worshipped became his wife.”

“He had help from Aphrodite. Looking back she’s sorry she did. All the gods have retreated from the idol worship they used to foster. When will regulators do the same?”

“They feel their banking system is safer when credits have top ratings—even though they lack the reserves to cover mistakes. The truth is closer to the opposite. When banking systems run low on reserves, even top-rated credits can become unsafe.”

“Reserves should be geared to the scope for surprise, not to risk itself,” said Pandora. “Most casinos take risky wagers all the time without harm. Their main concerns are attendance and the depth of pockets of those who attend. The analogous concern for banks is the phase of the credit cycle.”

“Regulators presume that safer bets have much lower variance than risky bets. That’s true when the risks are known. However, estimates of high safety tend to be highly uncertain. Additional buffers are needed to cover that.”

“Making rating agencies formally indicate their uncertainty might help. At the very least it will remind regulators and the public how fine credit gradations are, and that no one really knows what grade is deserved.”

“The rating agencies might not like that,” said Prometheus. “They profit from the aura of certainty. The regulators might like it even less. A lot of highly rated but uncertain credits are sovereign or sovereign-backed. They want that debt rolled over cheaply.”

“Acknowledging uncertainty needn’t undermine all confidence. Indeed, it might better justify the confidence we have. As the Ising-type model shows, a little deference can breed an unduly rigid consensus. Smoother adjustment might help the system weather big shocks.”